



Variational Bayesian Learning Theory

SHINICHI NAKAJIMA

KAZUHO WATANABE

MASASHI SUGIYAMA

Variational Bayesian Learning Theory

Variational Bayesian learning is one of the most popular methods in machine learning. Designed for researchers and graduate students in machine learning, this book summarizes recent developments in the nonasymptotic and asymptotic theory of variational Bayesian learning and suggests how this theory can be applied in practice.

The authors begin by developing a basic framework with a focus on conjugacy, which enables the reader to derive tractable algorithms. Next, it summarizes nonasymptotic theory, which, although limited in application to bilinear models, precisely describes the behavior of the variational Bayesian solution and reveals its sparsity-inducing mechanism. Finally, the text summarizes asymptotic theory, which reveals phase transition phenomena depending on the prior setting, thus providing suggestions on how to set hyperparameters for particular purposes. Detailed derivations allow readers to follow along without prior knowledge of the mathematical techniques specific to Bayesian learning.

SHINICHI NAKAJIMA is a senior researcher at Technische Universität Berlin. His research interests include the theory and applications of machine learning, and he has published papers at numerous conferences and in journals such as *The Journal of Machine Learning Research*, *The Machine Learning Journal*, *Neural Computation*, and *IEEE Transactions on Signal Processing*. He currently serves as an area chair for Neural Information Processing Systems (NIPS) and an action editor for Digital Signal Processing.

KAZUHO WATANABE is an associate professor at Toyohashi University of Technology. His research interests include statistical machine learning and information theory, and he has published papers at numerous conferences and in journals such as *The Journal of Machine Learning Research*, *The Machine Learning Journal*, *IEEE Transactions on Information Theory*, and *IEEE Transactions on Neural Networks and Learning Systems*.

MASASHI SUGIYAMA is the director of the RIKEN Center for Advanced Intelligence Project and professor of Complexity Science and Engineering at the University of Tokyo. His research interests include the theory, algorithms, and applications of machine learning. He has written several books on machine learning, including *Density Ratio Estimation in Machine Learning*. He served as program cochair and general cochair of the NIPS conference in 2015 and 2016, respectively, and received the Japan Academy Medal in 2017.

Variational Bayesian Learning Theory

SHINICHI NAKAJIMA

Technische Universität Berlin

KAZUHO WATANABE

Toyohashi University of Technology

MASASHI SUGIYAMA

University of Tokyo



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE

UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107076150

DOI: 10.1017/9781139879354

© Shinichi Nakajima, Kazuho Watanabe, and Masashi Sugiyama 2019

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2019

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Nakajima, Shinichi, author. | Watanabe, Kazuho, author. | Sugiyama,
Masashi, 1974- author.

Title: Variational Bayesian learning theory / Shinichi Nakajima (Technische
Universität Berlin), Kazuho Watanabe (Toyohashi University of
Technology), Masashi Sugiyama (University of Tokyo).

Description: Cambridge ; New York, NY : Cambridge University Press, 2019. |
Includes bibliographical references and index.

Identifiers: LCCN 2019005983 | ISBN 9781107076150 (hardback : alk. paper) |
ISBN 9781107430761 (pbk. : alk. paper)

Subjects: LCSH: Bayesian field theory. | Probabilities.

Classification: LCC QC174.85.B38 N35 2019 | DDC 519.2/33–dc23

LC record available at <https://lccn.loc.gov/2019005983>

ISBN 978-1-107-07615-0 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy
of URLs for external or third-party internet websites referred to in this publication
and does not guarantee that any content on such websites is, or will remain,
accurate or appropriate.

Contents

<i>Preface</i>	<i>page</i> ix
<i>Nomenclature</i>	xii
Part I Formulation	1
1 Bayesian Learning	3
1.1 Framework	3
1.2 Computation	10
2 Variational Bayesian Learning	39
2.1 Framework	39
2.2 Other Approximation Methods	51
Part II Algorithm	61
3 VB Algorithm for Multilinear Models	63
3.1 Matrix Factorization	63
3.2 Matrix Factorization with Missing Entries	74
3.3 Tensor Factorization	80
3.4 Low-Rank Subspace Clustering	87
3.5 Sparse Additive Matrix Factorization	93
4 VB Algorithm for Latent Variable Models	103
4.1 Finite Mixture Models	103
4.2 Other Latent Variable Models	115
5 VB Algorithm under No Conjugacy	132
5.1 Logistic Regression	132
5.2 Sparsity-Inducing Prior	135
5.3 Unified Approach by Local VB Bounds	137

Part III Nonasymptotic Theory	147
6 Global VB Solution of Fully Observed Matrix Factorization	149
6.1 Problem Description	150
6.2 Conditions for VB Solutions	152
6.3 Irrelevant Degrees of Freedom	153
6.4 Proof of Theorem 6.4	157
6.5 Problem Decomposition	160
6.6 Analytic Form of Global VB Solution	162
6.7 Proofs of Theorem 6.7 and Corollary 6.8	163
6.8 Analytic Form of Global Empirical VB Solution	171
6.9 Proof of Theorem 6.13	173
6.10 Summary of Intermediate Results	180
7 Model-Induced Regularization and Sparsity Inducing Mechanism	184
7.1 VB Solutions for Special Cases	184
7.2 Posteriors and Estimators in a One-Dimensional Case	187
7.3 Model-Induced Regularization	195
7.4 Phase Transition in VB Learning	202
7.5 Factorization as ARD Model	204
8 Performance Analysis of VB Matrix Factorization	205
8.1 Objective Function for Noise Variance Estimation	205
8.2 Bounds of Noise Variance Estimator	207
8.3 Proofs of Theorem 8.2 and Corollary 8.3	209
8.4 Performance Analysis	214
8.5 Numerical Verification	228
8.6 Comparison with Laplace Approximation	230
8.7 Optimality in Large-Scale Limit	232
9 Global Solver for Matrix Factorization	236
9.1 Global VB Solver for Fully Observed MF	236
9.2 Global EVB Solver for Fully Observed MF	238
9.3 Empirical Comparison with the Standard VB Algorithm	242
9.4 Extension to Nonconjugate MF with Missing Entries	247
10 Global Solver for Low-Rank Subspace Clustering	255
10.1 Problem Description	255
10.2 Conditions for VB Solutions	258
10.3 Irrelevant Degrees of Freedom	259
10.4 Proof of Theorem 10.2	259

10.5	Exact Global VB Solver (EGVBS)	264
10.6	Approximate Global VB Solver (AGVBS)	267
10.7	Proof of Theorem 10.7	270
10.8	Empirical Evaluation	274
11	Efficient Solver for Sparse Additive Matrix Factorization	279
11.1	Problem Description	279
11.2	Efficient Algorithm for SAMF	282
11.3	Experimental Results	284
12	MAP and Partially Bayesian Learning	294
12.1	Theoretical Analysis in Fully Observed MF	295
12.2	More General Cases	329
12.3	Experimental Results	332
Part IV Asymptotic Theory		339
13	Asymptotic Learning Theory	341
13.1	Statistical Learning Machines	341
13.2	Basic Tools for Asymptotic Analysis	344
13.3	Target Quantities	346
13.4	Asymptotic Learning Theory for Regular Models	351
13.5	Asymptotic Learning Theory for Singular Models	366
13.6	Asymptotic Learning Theory for VB Learning	382
14	Asymptotic VB Theory of Reduced Rank Regression	385
14.1	Reduced Rank Regression	385
14.2	Generalization Properties	396
14.3	Insights into VB Learning	426
15	Asymptotic VB Theory of Mixture Models	429
15.1	Basic Lemmas	429
15.2	Mixture of Gaussians	434
15.3	Mixture of Exponential Family Distributions	443
15.4	Mixture of Bernoulli with Deterministic Components	451
16	Asymptotic VB Theory of Other Latent Variable Models	455
16.1	Bayesian Networks	455
16.2	Hidden Markov Models	461
16.3	Probabilistic Context-Free Grammar	466
16.4	Latent Dirichlet Allocation	470

17 Unified Theory for Latent Variable Models	500
17.1 Local Latent Variable Model	500
17.2 Asymptotic Upper-Bound for VB Free Energy	504
17.3 Example: Average VB Free Energy of Gaussian Mixture Model	507
17.4 Free Energy and Generalization Error	511
17.5 Relation to Other Analyses	513
<i>Appendix A James–Stein Estimator</i>	516
<i>Appendix B Metric in Parameter Space</i>	520
<i>Appendix C Detailed Description of Overlap Method</i>	525
<i>Appendix D Optimality of Bayesian Learning</i>	527
<i>Bibliography</i>	529
<i>Subject Index</i>	540

Preface

Bayesian learning is a statistical inference method that provides estimators and other quantities computed from the *posterior distribution*—the conditional distribution of unknown variables given observed variables. Compared with *point estimation* methods such as maximum likelihood (ML) estimation and maximum a posteriori (MAP) learning, Bayesian learning has the following advantages:

- Theoretically optimal.

The posterior distribution is what we can obtain best about the unknown variables from observation. Therefore, Bayesian learning provides most accurate predictions, provided that the assumed model is appropriate.

- Uncertainty information is available.

Sharpness of the posterior distribution indicates the reliability of estimators. The credible interval, which can be computed from the posterior distribution, provides probabilistic bounds of unknown variables.

- Model selection and hyperparameter estimation can be performed in a single framework.

The marginal likelihood can be used as a criterion to evaluate how well a statistical model (which is typically a combination of model and prior distributions) fits the observed data, taking account of the flexibility of the model as a penalty.

- Less prone to overfitting.

It was theoretically proven that Bayesian learning overfits the observation noise less than MAP learning.

On the other hand, Bayesian learning has a critical drawback—computing the posterior distribution is computationally hard in many practical models. This is because Bayesian learning requires *expectation* operations or integral computations, which cannot be analytically performed except for simple cases.

Accordingly, various approximation methods, including deterministic and sampling methods, have been proposed.

Variational Bayesian (VB) learning is one of the most popular deterministic approximation methods to Bayesian learning. VB learning aims to find the closest distribution to the Bayes posterior under some constraints, which are designed so that the expectation operation is tractable. The simplest and most popular approach is the *mean field approximation* where the approximate posterior is sought in the space of *decomposable* distributions, i.e., groups of unknown variables are forced to be independent of each other. In many practical models, Bayesian learning is intractable *jointly* for all unknown parameters, while it is tractable if the dependence between groups of parameters is ignored. Such a case often happens because many practical models have been constructed by combining simple models in which Bayesian learning is analytically tractable. This property is called *conditional conjugacy*, and makes VB learning computationally tractable.

Since its development, VB learning has shown good performance in many applications. Its good aspects and downsides have been empirically observed and qualitatively discussed. Some of those aspects seem inherited from full Bayesian learning, while some others seem to be artifacts by forced independence constraints. We have dedicated ourselves to theoretically clarifying the behavior of VB learning quantitatively, which is the main topic of this book.

This book starts from the formulation of Bayesian learning methods. In Part I, we introduce Bayesian learning and VB learning, emphasizing how conjugacy and conditional conjugacy make the computation tractable. We also briefly introduce other approximation methods and relate them to VB learning. In Part II, we derive algorithms of VB learning for popular statistical models, on which theoretical analysis will be conducted in the subsequent parts.

We categorize the theory of VB learning into two parts, and exhibit them separately. Part III focuses on *nonasymptotic* theory, where we do not assume the availability of a large number of samples. This analysis so far has been applied only to a class of *bilinear* models, but we can make detailed discussions including analytic forms of global solutions and theoretical performance guarantees. On the other hand, Part IV focuses on asymptotic theory, where the number of observed samples is assumed to be large. This approach has been applied to a broad range of statistical models, and successfully elucidated the *phase transition* phenomenon of VB learning. As a practical outcome, this analysis provides a guideline on how to set hyperparameters for different purposes.

Recently, a lot of variations of VB learning have been proposed, e.g., more accurate inference methods beyond the mean field approximation, stochastic gradient optimization for big data analysis, and sampling based update rules for automatic (black-box) inference to cope with general nonconjugate likelihoods including deep neural networks. Although we briefly introduce some of those recent works in Part I, they are not in the central scope of this book. We rather focus on the simplest mean field approximation, of which the behavior has been clarified quantitatively by theory.

This book was completed under the support by many people. Shinichi Nakajima deeply thanks Professor Klaus-Robert Müller and the members in Machine Learning Group in Technische Universität Berlin for their direct and indirect support during the period of book writing. Special thanks go to Sergej Dogadov, Hannah Marienwald, Ludwig Winkler, Dr. Nico Gönitz, and Dr. Pan Kessel, who reviewed chapters of earlier versions, found errors and typos, provided suggestions to improve the presentation, and kept encouraging him in proceeding book writing. The authors also thank Lauren Cowles and her team in Cambridge University Press, as well as all other staff members who contributed to the book production process, for their help, as well as their patience on the delays in our manuscript preparation. Lauren Cowles, Clare Dennison, Adam Kratoska, and Amy He have coordinated the project since its proposal, and Harsha Vardhanan in SPi Global has managed the copy-editing process with Andy Saff.

The book writing project was partially supported by the following organizations: the German Research Foundation (GRK 1589/1) by the Federal Ministry of Education and Research (BMBF) under the Berlin Big Data Center project (Phase 1: FKZ 01IS14013A and Phase 2: FKz 01IS18025A), the Japan Society for the Promotion of Science (15K16050), and the International Research Center for Neurointelligence (WPI-IRCN) at The University of Tokyo Institutes for Advanced Study.

Nomenclature

$a, b, c, \dots, A, B, C, \dots$: Scalars.
$\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$ (bold-faced small letters)	: Vectors.
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots$ (bold-faced capital letters)	: Matrices.
$\mathcal{A}, \mathcal{B}, \mathcal{C}, \dots$ (calligraphic capital letters)	: Tensors or sets.
$(\cdot)_{l,m}$: (l, m) th element of a matrix.
\top	: Transpose of a matrix or vector.
$\text{tr}(\cdot)$: Trace of a matrix.
$\det(\cdot)$: Determinant of a matrix.
\odot	: Hadamard (elementwise) product.
\otimes	: Kronecker product.
\times_n	: n -mode tensor product.
$ \cdot $: Absolute value of a scalar. It applies element-wise for a vector or matrix.
$\text{sign}(\cdot)$: Sign operator such that $\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ -1 & \text{otherwise.} \end{cases}$ It applies elementwise for a vector or matrix.
$\{\dots\}$: Set consisting of specified entities.
$\{\dots\}^D$: D fold Cartesian product, i.e., $\mathbb{X}^D \equiv \{(x_1, \dots, x_D)^\top; x_d \in \mathbb{X} \text{ for } d = 1, \dots, D\}.$
$\#(\cdot)$: Cardinality (the number of entities) of a set.
\mathbb{R}	: The set of all real numbers.
\mathbb{R}_+	: The set of all nonnegative real numbers.
\mathbb{R}_{++}	: The set of all positive real numbers.
\mathbb{R}^D	: The set of all D -dimensional real (column) vectors.

$[\cdot, \cdot]$: The set of real numbers in a range, i.e., $[l, u] = \{x \in \mathbb{R}; l \leq x \leq u\}.$
$[\cdot, \cdot]^D$: The set of D -dimensional real vectors whose entries are in a range, i.e., $[l, u]^D \equiv \{\mathbf{x} \in \mathbb{R}^D; l \leq x_d \leq u \text{ for } d = 1, \dots, D\}.$
$\mathbb{R}^{L \times M}$: The set of all $L \times M$ real matrices.
$\mathbb{R}^{M_1 \times M_2 \times \dots \times M_N}$: The set of all $M_1 \times M_2 \times \dots \times M_N$ real tensors.
\mathbb{I}	: The set of all integers.
\mathbb{I}_{++}	: The set of all positive integers.
\mathbb{C}	: The set of all complex numbers.
\mathbb{S}^D	: The set of all $D \times D$ symmetric matrices.
\mathbb{S}_+^D	: The set of all $D \times D$ positive semidefinite matrices.
\mathbb{S}_{++}^D	: The set of all $D \times D$ positive definite matrices.
\mathbb{D}^D	: The set of all $D \times D$ diagonal matrices.
\mathbb{D}_+^D	: The set of all $D \times D$ positive semidefinite diagonal matrices.
\mathbb{D}_{++}^D	: The set of all $D \times D$ positive definite diagonal matrices.
\mathbb{H}_N^{K-1}	: The set of all possible histograms for N samples and K categories, i.e., $\mathbb{H}_N^{K-1} \equiv \{\mathbf{x} \in \{0, \dots, N\}^K; \sum_{k=1}^K x_k = N\}.$
Δ^{K-1}	: The standard $(K-1)$ -simplex, i.e., $\Delta^{K-1} \equiv \{\boldsymbol{\theta} \in [0, 1]^K; \sum_{k=1}^K \theta_k = 1\}.$
$(\mathbf{a}_1, \dots, \mathbf{a}_M)$: Column vectors of A , i.e., $A = (\mathbf{a}_1, \dots, \mathbf{a}_M) \in \mathbb{R}^{L \times M}.$
$(\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_L)$: Row vectors of A , i.e., $A = (\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_L)^\top \in \mathbb{R}^{L \times M}.$
Diag (\cdot)	: Diagonal matrix with specified diagonal elements, i.e., $(\mathbf{Diag}(\mathbf{x}))_{l,m} = \begin{cases} x_l & \text{if } l = m, \\ 0 & \text{otherwise.} \end{cases}$
diag (\cdot)	: Column vector consisting of the diagonal entries of a matrix, i.e., $(\mathbf{diag}(X))_l = X_{l,l}.$
vec (\cdot)	: Vectorization operator concatenating all column vectors of a matrix into a long column vector, i.e., $\mathbf{vec}(A) = (\mathbf{a}_1^\top, \dots, \mathbf{a}_M^\top)^\top \in \mathbb{R}^{LM}$ for a matrix $A = (\mathbf{a}_1, \dots, \mathbf{a}_M) \in \mathbb{R}^{L \times M}.$
I_D	: D -dimensional ($D \times D$) identity matrix.
$\boldsymbol{\Gamma}$: A diagonal matrix.
$\boldsymbol{\Omega}$: An orthogonal matrix.
\mathbf{e}_k	: One of K expression, i.e., $\mathbf{e}_k = (\underbrace{0, \dots, 0}_{K}, \underbrace{1}_{k^{\text{th}}}, \underbrace{0, \dots, 0}_{K})^\top \in \{0, 1\}^K.$
$\mathbf{1}_K$: K -dimensional vector with all elements equal to one, i.e., $\mathbf{e}_k = (\underbrace{1, \dots, 1}_K)^\top.$

$\text{Gauss}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: D -dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
$\text{MGauss}_{D_1, D_2}(\mathbf{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$: $D_1 \times D_2$ dimensional matrix variate Gaussian distribution with mean \mathbf{M} and covariance $\boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}$.
$\text{Gamma}(\alpha, \beta)$: Gamma distribution with shape parameter α and scale parameter β .
$\text{InvGamma}(\alpha, \beta)$: Inverse-Gamma distribution with shape parameter α and scale parameter β .
$\text{Wishart}_D(\mathbf{V}, \nu)$: D -dimensional Wishart distribution with scale matrix \mathbf{V} and degree of freedom ν .
$\text{InvWishart}_D(\mathbf{V}, \nu)$: D -dimensional inverse-Wishart distribution with scale matrix \mathbf{V} and degree of freedom ν .
$\text{Multinomial}(\boldsymbol{\theta}, N)$: Multinomial distribution with event probabilities $\boldsymbol{\theta}$ and number of trials N .
$\text{Dirichlet}(\boldsymbol{\phi})$: Dirichlet distribution with concentration parameters $\boldsymbol{\phi}$.
$\text{Prob}(\cdot)$: Probability of an event.
$p(\cdot), q(\cdot)$: Probability distribution (probability mass function for discrete random variables, and probability density function for continuous random variables). Typically p is used for a model distribution and q is used for the true distribution.
$r(\cdot)$: A trial distribution (a variable of a functional) for approximation.
$\langle f(\mathbf{x}) \rangle_{p(\mathbf{x})}$: Expectation value of $f(\mathbf{x})$ over distribution $p(\mathbf{x})$, i.e., $\langle f(\mathbf{x}) \rangle_{p(\mathbf{x})} \equiv \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$
$\widehat{\cdot}$: Estimator for an unknown variable, e.g., $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{A}}$ are estimators for a vector \mathbf{x} and a matrix \mathbf{A} , respectively.
$\text{Mean}(\cdot)$: Mean of a random variable.
$\text{Var}(\cdot)$: Variance of a random variable.
$\text{Cov}(\cdot)$: Covariance of a random variable.
$\text{KL}(\cdot \ \cdot)$: Kullback–Leibler divergence between distributions, i.e., $\text{KL}(p \ q) \equiv \left\langle \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \right\rangle_{p(\mathbf{x})}.$
$\delta(\boldsymbol{\mu}; \widehat{\boldsymbol{\mu}})$: Dirac delta function located at $\widehat{\boldsymbol{\mu}}$. It also denotes its approximation (called Pseudo-delta function) with its entropy finite.
GE	: Generalization error.
TE	: Training error.
F	: Free energy.

$O(f(N))$: A function such that $\limsup_{N \rightarrow \infty} O(f(N))/f(N) < \infty$.
$o(f(N))$: A function such that $\lim_{N \rightarrow \infty} o(f(N))/f(N) = 0$.
$\Omega(f(N))$: A function such that $\liminf_{N \rightarrow \infty} \Omega(f(N))/f(N) > 0$
$\omega(f(N))$: A function such that $\lim_{N \rightarrow \infty} \omega(f(N))/f(N) = \infty$.
$\Theta(f(N))$: A function such that $\limsup_{N \rightarrow \infty} \Theta(f(N))/f(N) < \infty$ and $\liminf_{N \rightarrow \infty} \Theta(f(N))/f(N) > 0$.
$O_p(f(N))$: A function such that $\limsup_{N \rightarrow \infty} O_p(f(N))/f(N) < \infty$ in probability.
$o_p(f(N))$: A function such that $\lim_{N \rightarrow \infty} o_p(f(N))/f(N) = 0$ in probability.
$\Omega_p(f(N))$: A function such that $\liminf_{N \rightarrow \infty} \Omega_p(f(N))/f(N) > 0$ in probability
$\omega_p(f(N))$: A function such that $\lim_{N \rightarrow \infty} \omega_p(f(N))/f(N) = \infty$ in probability.
$\Theta_p(f(N))$: A function such that $\limsup_{N \rightarrow \infty} \Theta_p(f(N))/f(N) < \infty$ and $\liminf_{N \rightarrow \infty} \Theta_p(f(N))/f(N) > 0$ in probability.

Part I

Formulation

1

Bayesian Learning

Bayesian learning is an inference method based on the fundamental law of probability, called the Bayes theorem. In this first chapter, we introduce the framework of Bayesian learning with simple examples where Bayesian learning can be performed analytically.

1.1 Framework

Bayesian learning considers the following situation. We have observed a set \mathcal{D} of data, which are subject to a *conditional distribution* $p(\mathcal{D}|\mathbf{w})$, called the *model distribution*, of the data given unknown *model parameter* \mathbf{w} . Although the value of \mathbf{w} is unknown, vague information on \mathbf{w} is provided as a *prior distribution* $p(\mathbf{w})$. The conditional distribution $p(\mathcal{D}|\mathbf{w})$ is also called the *model likelihood* when it is seen as a function of the unknown parameter \mathbf{w} .

1.1.1 Bayes Theorem and Bayes Posterior

Bayesian learning is based on the following basic factorization property of the *joint distribution* $p(\mathcal{D}, \mathbf{w})$:

$$\underbrace{p(\mathbf{w}|\mathcal{D})}_{\text{posterior}} \underbrace{p(\mathcal{D})}_{\text{marginal}} = \underbrace{p(\mathcal{D}, \mathbf{w})}_{\text{joint}} = \underbrace{p(\mathcal{D}|\mathbf{w})}_{\text{likelihood}} \underbrace{p(\mathbf{w})}_{\text{prior}}, \quad (1.1)$$

where the marginal distribution is given by

$$p(\mathcal{D}) = \int_{\mathcal{W}} p(\mathcal{D}, \mathbf{w}) d\mathbf{w} = \int_{\mathcal{W}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}) d\mathbf{w}. \quad (1.2)$$

Here, the integration is performed in the domain \mathcal{W} of the parameter \mathbf{w} . Note that, if the domain \mathcal{W} is discrete, integration should be replaced with

summation, i.e., for any function $f(\mathbf{w})$,

$$\int_{\mathcal{W}} f(\mathbf{w}) d\mathbf{w} \rightarrow \sum_{\mathbf{w}' \in \mathcal{W}} f(\mathbf{w}').$$

The *posterior distribution*, the distribution of the unknown parameter \mathbf{w} given the observed data set \mathcal{D} , is derived by dividing both sides of Eq. (1.1) by the marginal distribution $p(\mathcal{D})$:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}, \mathbf{w})}{p(\mathcal{D})} \propto p(\mathcal{D}, \mathbf{w}). \quad (1.3)$$

Here, we emphasized that the posterior distribution is proportional to the joint distribution $p(\mathcal{D}, \mathbf{w})$ because the marginal distribution $p(\mathcal{D})$ is a constant (as a function of \mathbf{w}). In other words, the joint distribution is an *unnormalized posterior distribution*. Eq. (1.3) is called the *Bayes theorem*, and the posterior distribution computed exactly by Eq. (1.3) is called the *Bayes posterior* when we distinguish it from its approximations.

Example 1.1 (Parametric density estimation) Assume that the observed data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ consist of N *independent and identically distributed* (i.i.d.) samples from the model distribution $p(\mathbf{x}|\mathbf{w})$. Then, the model likelihood is given by $p(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w})$, and therefore, the posterior distribution is given by

$$p(\mathbf{w}|\mathcal{D}) = \frac{\prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) p(\mathbf{w})}{\int \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w}} \propto \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) p(\mathbf{w}).$$

Example 1.2 (Parametric regression) Assume that the observed data $\mathcal{D} = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})\}$ consist of N i.i.d. input–output pairs from the model distribution $p(\mathbf{x}, \mathbf{y}|\mathbf{w}) = p(\mathbf{y}|\mathbf{x}, \mathbf{w})p(\mathbf{x})$. Then, the likelihood function is given by $p(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^N p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \mathbf{w})p(\mathbf{x}^{(n)})$, and therefore, the posterior distribution is given by

$$p(\mathbf{w}|\mathcal{D}) = \frac{\prod_{n=1}^N p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \mathbf{w}) p(\mathbf{w})}{\int \prod_{n=1}^N p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}} \propto \prod_{n=1}^N p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \mathbf{w}) p(\mathbf{w}).$$

Note that the input distribution $p(\mathbf{x})$ does not affect the posterior, and accordingly is often ignored in practice.

1.1.2 Maximum A Posteriori Learning

Since the joint distribution $p(\mathcal{D}, \mathbf{w})$ is just the product of the likelihood function and the prior distribution (see Eq. (1.1)), it is usually easy to

compute. Therefore, it is relatively easy to perform *maximum a posteriori (MAP) learning*, where the parameters are point-estimated so that the posterior probability is maximized, i.e.,

$$\widehat{\mathbf{w}}^{\text{MAP}} = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{w}|\mathcal{D}) = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathcal{D}, \mathbf{w}). \quad (1.4)$$

MAP learning includes *maximum likelihood (ML) learning*,

$$\widehat{\mathbf{w}}^{\text{ML}} = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathcal{D}|\mathbf{w}), \quad (1.5)$$

as a special case with the flat prior $p(\mathbf{w}) \propto 1$.

1.1.3 Bayesian Learning

On the other hand, *Bayesian learning* requires integration of the joint distribution with respect to the parameter \mathbf{w} , which is often computationally hard. More specifically, performing Bayesian learning means computing at least one of the following quantities:

Marginal likelihood (zeroth moment)

$$p(\mathcal{D}) = \int p(\mathcal{D}, \mathbf{w}) d\mathbf{w}. \quad (1.6)$$

This quantity has been already introduced in Eq. (1.2) as the normalization factor of the posterior distribution. As seen in Section 1.1.5 and subsequent sections, the marginal likelihood plays an important role in model selection and hyperparameter estimation.

Posterior mean (first moment)

$$\widehat{\mathbf{w}} = \langle \mathbf{w} \rangle_{p(\mathbf{w}|\mathcal{D})} = \frac{1}{p(\mathcal{D})} \int \mathbf{w} \cdot p(\mathcal{D}, \mathbf{w}) d\mathbf{w}, \quad (1.7)$$

where $\langle \cdot \rangle_p$ denotes the expectation value over the distribution p , i.e., $\langle \cdot \rangle_{p(\mathbf{w})} = \int \cdot p(\mathbf{w}) d\mathbf{w}$. This quantity is also called the *Bayesian estimator*. The Bayesian estimator or the model distribution with the Bayesian estimator plugged in (see the plug-in predictive distribution (1.10)) can be the final output of Bayesian learning.

Posterior covariance (second moment)

$$\widehat{\Sigma}_{\mathbf{w}} = \left\langle (\mathbf{w} - \widehat{\mathbf{w}})(\mathbf{w} - \widehat{\mathbf{w}})^{\top} \right\rangle_{p(\mathbf{w}|\mathcal{D})} = \frac{1}{p(\mathcal{D})} \int (\mathbf{w} - \widehat{\mathbf{w}})(\mathbf{w} - \widehat{\mathbf{w}})^{\top} p(\mathcal{D}, \mathbf{w}) d\mathbf{w}, \quad (1.8)$$

where \top denotes the transpose of a matrix or vector. This quantity provides the credibility information, and is used to assess the confidence level of the Bayesian estimator.

Predictive distribution (expectation of model distribution)

$$p(\mathcal{D}^{\text{new}}|\mathcal{D}) = \langle p(\mathcal{D}^{\text{new}}|\mathbf{w}) \rangle_{p(\mathbf{w}|\mathcal{D})} = \frac{1}{p(\mathcal{D})} \int p(\mathcal{D}^{\text{new}}|\mathbf{w})p(\mathcal{D}, \mathbf{w})d\mathbf{w}, \quad (1.9)$$

where $p(\mathcal{D}^{\text{new}}|\mathbf{w})$ denotes the model distribution on *unobserved* new data \mathcal{D}^{new} . In the i.i.d. case such as Examples 1.1 and 1.2, it is sufficient to compute the predictive distribution for a single new sample $\mathcal{D}^{\text{new}} = \{\mathbf{x}\}$.

Note that each of the four quantities (1.6) through (1.9) requires to compute the expectation of some function $f(\mathbf{w})$ over the unnormalized posterior distribution $p(\mathcal{D}, \mathbf{w})$ on \mathbf{w} , i.e., $\int f(\mathbf{w})p(\mathcal{D}, \mathbf{w})d\mathbf{w}$. Specifically, the marginal likelihood, the posterior mean, and the posterior covariance are the zeroth, the first, and the second moments of the unnormalized posterior distribution, respectively. The expectation is analytically intractable except for some simple cases, and numerical computation is also hard when the dimensionality of the unknown parameter \mathbf{w} is high. This is the main bottleneck of Bayesian learning, with which many approximation methods have been developed to cope.

It hardly happens that the first moment (1.7) or the second moment (1.8) are computationally tractable but the zeroth moment (1.6) is not. Accordingly, we can say that performing Bayesian learning on the parameter \mathbf{w} amounts to obtaining the *normalized* posterior distribution $p(\mathbf{w}|\mathcal{D})$. It sometimes happens that computing the predictive distribution (1.9) is still intractable even if the zeroth, the first, and the second moments can be computed based on some approximation. In such a case, the model distribution with the Bayesian estimator plugged in, called the *plug-in predictive distribution*,

$$p(\mathcal{D}^{\text{new}}|\widehat{\mathbf{w}}), \quad (1.10)$$

is used for prediction in practice.

1.1.4 Latent Variables

So far, we introduced the observed data set \mathcal{D} as a known variable, and the model parameter \mathbf{w} as an unknown variable. In practice, more varieties of known and unknown variables can be involved.

Some probabilistic models have *latent variables* (or *hidden variables*) \mathbf{z} , which can be involved in the original model, or additionally introduced for

computational reasons. They are typically attributed to each of the observed samples, and therefore have large degrees of freedom. However, they are just additional unknown variables, and there is no reason in inference to distinguish them from the model parameters \mathbf{w} .¹ The joint posterior over the parameters and the latent variables is given by Eq. (1.3) with \mathbf{w} and $p(\mathbf{w})$ replaced with $\bar{\mathbf{w}} = (\mathbf{w}, \mathbf{z})$ and $p(\bar{\mathbf{w}}) = p(\mathbf{z}|\mathbf{w})p(\mathbf{w})$, respectively.

Example 1.3 (Mixture models) A mixture model is often used for parametric density estimation (Example 1.1). The model distribution is given by

$$p(\mathbf{x}|\mathbf{w}) = \sum_{k=1}^K \alpha_k p(\mathbf{x}|\tau_k), \quad (1.11)$$

where $\mathbf{w} = \{\alpha_k, \tau_k; \alpha_k \geq 0, \sum_{k=1}^K \alpha_k = 1\}_{k=1}^K$ is the unknown parameters. The mixture model (1.11) is the weighted sum of K distributions, each of which is parameterized by the component parameter τ_k . The domain of the *mixing weights* $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\top$, also called as the *mixture coefficients*, forms the *standard $(K - 1)$ -simplex*, denoted by $\Delta^{K-1} \equiv \{\boldsymbol{\alpha} \in \mathbb{R}_+^K; \sum_{k=1}^K \alpha_k = 1\}$ (see Figure 1.1). Figure 1.2 shows an example of the mixture model with three one-dimensional Gaussian components.

The likelihood,

$$\begin{aligned} p(\mathcal{D}|\mathbf{w}) &= \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}), \\ &= \prod_{n=1}^N \left(\sum_{k=1}^K \alpha_k p(\mathbf{x}|\tau_k) \right), \end{aligned} \quad (1.12)$$

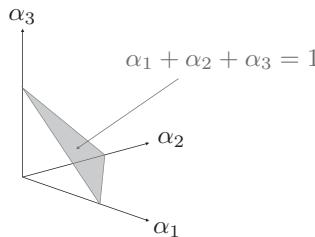


Figure 1.1 $(K - 1)$ -simplex, Δ^{K-1} , for $K = 3$.

¹ For this reason, the latent variables \mathbf{z} and the model parameters \mathbf{w} are also called *local latent variables* and *global latent variables*, respectively.

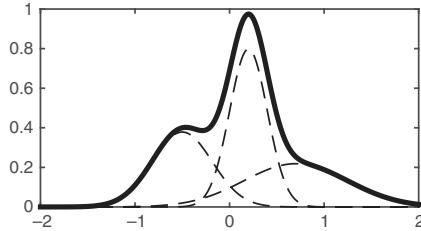


Figure 1.2 Gaussian mixture.

for N observed i.i.d. samples $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ has $O(K^N)$ terms, which makes even ML learning intractable. This intractability arises from the summation inside the multiplication in Eq. (1.12). By introducing latent variables, we can turn this summation into a multiplication, and make Eq. (1.12) tractable.

Assume that each sample \mathbf{x} belongs to a single component k , and is drawn from $p(\mathbf{x}|\tau_k)$. To describe the assignment, we introduce a latent variable $\mathbf{z} \in \mathcal{Z} \equiv \{\mathbf{e}_k\}_{k=1}^K$ associated with each observed sample \mathbf{x} , where $\mathbf{e}_k \in \{0, 1\}^K$ is the K -dimensional binary vector, called the *one-of- K representation*, with one at the k th entry and zeros at the other entries:

$$\mathbf{e}_k = (\underbrace{0, \dots, 0, \overbrace{1}^{k\text{th}}, 0, \dots, 0}_{K}, 0)^T.$$

Then, we have the following model:

$$p(\mathbf{x}, \mathbf{z}|\mathbf{w}) = p(\mathbf{x}|\mathbf{z}, \mathbf{w})p(\mathbf{z}|\mathbf{w}), \quad (1.13)$$

$$\text{where } p(\mathbf{x}|\mathbf{z}, \mathbf{w}) = \prod_{k=1}^K \{p(\mathbf{x}|\tau_k)\}^{z_k}, \quad p(\mathbf{z}|\mathbf{w}) = \prod_{k=1}^K \alpha_k^{z_k}.$$

The conditional distribution (1.13) on the observed variable \mathbf{x} and the latent variable \mathbf{z} given the parameter \mathbf{w} is called the *complete likelihood*.

Note that marginalizing the complete likelihood over the latent variable recovers the original mixture model:

$$p(\mathbf{x}|\mathbf{w}) = \int_{\mathcal{Z}} p(\mathbf{x}, \mathbf{z}|\mathbf{w}) d\mathbf{z} = \sum_{\mathbf{z} \in \{\mathbf{e}_k\}_{k=1}^K} \prod_{k=1}^K \{\alpha_k p(\mathbf{x}|\tau_k)\}^{z_k} = \sum_{k=1}^K \alpha_k p(\mathbf{x}|\tau_k).$$

This means that, if samples are generated from the model distribution (1.13), and only \mathbf{x} is recorded, the observed data follow the original mixture model (1.11).

In the literature, latent variables tend to be marginalized out even in MAP learning. For example, the *expectation-maximization (EM) algorithm* (Dempster et al., 1977), a popular MAP solver for latent variable models, seeks a (local) maximizer of the posterior distribution with the latent variables marginalized out, i.e.,

$$\widehat{\mathbf{w}}^{\text{EM}} = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{w}|\mathcal{D}) = \underset{\mathbf{w}}{\operatorname{argmax}} \int_{\mathcal{Z}} p(\mathcal{D}, \mathbf{w}, \mathbf{z}) d\mathbf{z}. \quad (1.14)$$

However, we can also maximize the posterior jointly over the parameters and the latent variables, i.e.,

$$(\widehat{\mathbf{w}}^{\text{MAP-hard}}, \widehat{\mathbf{z}}^{\text{MAP-hard}}) = \underset{\mathbf{w}, \mathbf{z}}{\operatorname{argmax}} p(\mathbf{w}, \mathbf{z}|\mathcal{D}) = \underset{\mathbf{w}, \mathbf{z}}{\operatorname{argmax}} p(\mathcal{D}, \mathbf{w}, \mathbf{z}). \quad (1.15)$$

For clustering based on the mixture model in Example 1.3, the EM algorithm (1.14) gives a *soft assignment*, where the expectation value $\widehat{\mathbf{z}}^{\text{EM}} \in \Delta^{K-1} \subset [0, 1]^K$ is substituted into the joint distribution $p(\mathcal{D}, \mathbf{w}, \mathbf{z})$, while the joint maximization (1.15) gives the *hard assignment*, where the optimal assignment $\widehat{\mathbf{z}}^{\text{MAP-hard}} \in \{\mathbf{e}_k\}_{k=1}^K \subset \{0, 1\}^K$ is looked for in the binary domain.

1.1.5 Empirical Bayesian Learning

In many practical cases, it is reasonable to use a prior distribution parameterized by *hyperparameters* $\boldsymbol{\kappa}$. The hyperparameters can be tuned by hand or based on some criterion outside the Bayesian framework. A popular method of the latter is the *cross validation*, where the hyperparameters are tuned so that an (preferably unbiased) estimator of the performance criterion is optimized. In such cases, the hyperparameters should be treated as *known* variables when Bayesian learning is performed.

On the other hand, the hyperparameters can be estimated within the Bayesian framework. In this case, there is again no reason to distinguish the hyperparameters from the other unknown variables (\mathbf{w}, \mathbf{z}) . The joint posterior over all unknown variables is given by Eq. (1.3) with \mathbf{w} and $p(\mathbf{w})$ replaced with $\overline{\mathbf{w}} = (\mathbf{w}, \boldsymbol{\kappa}, \mathbf{z})$ and $p(\overline{\mathbf{w}}) = p(\mathbf{z}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\kappa})p(\boldsymbol{\kappa})$, respectively, where $p(\boldsymbol{\kappa})$ is called a *hyperprior*. A popular approach, called *empirical Bayesian (EBayes) learning* (Efron and Morris, 1973), applies Bayesian learning on \mathbf{w} (and \mathbf{z}) and point-estimate $\boldsymbol{\kappa}$, i.e.,

$$\widehat{\boldsymbol{\kappa}}^{\text{EBayes}} = \underset{\boldsymbol{\kappa}}{\operatorname{argmax}} p(\mathcal{D}, \boldsymbol{\kappa}) = \underset{\boldsymbol{\kappa}}{\operatorname{argmax}} p(\mathcal{D}|\boldsymbol{\kappa})p(\boldsymbol{\kappa}),$$

$$\text{where } p(\mathcal{D}|\boldsymbol{\kappa}) = \int p(\mathcal{D}, \mathbf{w}, \mathbf{z}|\boldsymbol{\kappa}) d\mathbf{w} d\mathbf{z}.$$

Here the marginal likelihood $p(\mathcal{D}|\kappa)$ is seen as the likelihood of the hyperparameter κ , and MAP learning is performed by maximizing the joint distribution $p(\mathcal{D}, \kappa)$ of the observed data \mathcal{D} and the hyperparameter κ , which can be seen as an *unnormalized posterior distribution* of the hyperparameter. The hyperprior is often assumed to be flat: $p(\kappa) \propto 1$.

With an appropriate design of priors, empirical Bayesian learning combined with approximate Bayesian learning is often used for *automatic relevance determination (ARD)*, where irrelevant degrees of freedom of the statistical model are automatically pruned out. Explaining the ARD property of approximate Bayesian learning is one of the main topics of theoretical analysis in Parts III and IV.

1.2 Computation

Now, let us explain how Bayesian learning is performed in simple cases. We start from introducing *conjugacy*, an important notion in performing Bayesian learning.

1.2.1 Popular Distributions

Table 1.1 summarizes several distributions that are frequently used as a model distribution (or likelihood function) $p(\mathcal{D}|w)$ or a prior distribution $p(w)$ in Bayesian learning. The domain X of the random variable x and the domain W of the parameters w are shown in the table.

Some of the distributions in Table 1.1 have complicated function forms, involving Beta or Gamma functions. However, such complications are mostly in the *normalization constant*, and can often be ignored when it is sufficient to find the *shape* of a function. In Table 1.1, the normalization constant is separated by a dot, so that one can find the simple main part. As will be seen shortly, we often refer to the normalization constant when we need to perform integration of a function, which is in the same form as the main part of a popular distribution.

Below we summarize abbreviations of distributions:

$$\text{Gauss}_M(x; \mu, \Sigma) \equiv \frac{1}{(2\pi)^{M/2} \det(\Sigma)^{1/2}} \cdot \exp\left(-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu)\right), \quad (1.16)$$

$$\text{Gamma}(x; \alpha, \beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} \exp(-\beta x), \quad (1.17)$$

Table 1.1 *Popular distributions.* The following notation is used: \mathbb{R} : The set of all real numbers, \mathbb{R}_{++} : The set of all positive real numbers, \mathbb{I}_{++} : The set of all positive integers, \mathbb{S}_{++}^M : The set of all $M \times M$ positive definite matrices, $\mathbb{H}_N^{K-1} \equiv \{\mathbf{x} \in \{0, \dots, N\}^K; \sum_{k=1}^K x_k = N\}$: The set of all possible histograms for N samples and K categories, $\Delta^{K-1} \equiv \{\boldsymbol{\theta} \in [0, 1]^K; \sum_{k=1}^K \theta_k = 1\}$: The standard $(K - 1)$ -simplex, $\det(\cdot)$: Determinant of matrix, $\mathcal{B}(y, z) \equiv \int_0^1 t^{y-1} (1-t)^{z-1} dt$: Beta function, $\Gamma(y) \equiv \int_0^\infty t^{y-1} \exp(-t) dt$: Gamma function, and $\Gamma_M(y) \equiv \int_{T \in \mathbb{S}_{++}^M} \det(T)^{y-(M+1)/2} \exp(-\text{tr}(T)) dT$: Multivariate Gamma function.

Probability distribution	$p(\mathbf{x} \mathbf{w})$	$\mathbf{x} \in \mathcal{X}$	$\mathbf{w} \in \mathcal{W}$
Isotropic Gaussian	$\text{Gauss}_M(\mathbf{x}; \boldsymbol{\mu}, \sigma^2 \mathbf{I}_M) \equiv \frac{1}{(2\pi\sigma^2)^{M/2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\ \mathbf{x} - \boldsymbol{\mu}\ ^2\right)$	$\mathbf{x} \in \mathbb{R}^M$	$\boldsymbol{\mu} \in \mathbb{R}^M, \sigma^2 > 0$
Gaussian	$\text{Gauss}_M(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$	$\mathbf{x} \in \mathbb{R}^M$	$\boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\Sigma} \in \mathbb{S}_{++}^M$
Gamma	$\text{Gamma}(x; \alpha, \beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} \exp(-\beta x)$	$x > 0$	$\alpha > 0, \beta > 0$
Wishart	$\text{Wishart}_M(\mathbf{X}; \mathbf{V}, v) \equiv \frac{1}{(2^v \mathbf{V})^{M/2} \Gamma_M(\frac{v}{2})} \cdot \det(\mathbf{X})^{\frac{v-M-1}{2}} \exp\left(-\frac{\text{tr}(\mathbf{V}^{-1} \mathbf{X})}{2}\right)$	$\mathbf{X} \in \mathbb{S}_{++}^M$	$\mathbf{V} \in \mathbb{S}_{++}^M, v > M - 1$
Bernoulli	$\text{Binomial}_1(x; \theta) \equiv \theta^x (1-\theta)^{1-x}$	$x \in \{0, 1\}$	$\theta \in [0, 1]$
Binomial	$\text{Binomial}_N(x; \theta) \equiv \binom{N}{x} \theta^x (1-\theta)^{N-x}$	$x \in \{0, \dots, N\}$	$\theta \in [0, 1]$
Multinomial	$\text{Multinomial}_{K,N}(\mathbf{x}; \boldsymbol{\theta}) \equiv N! \cdot \prod_{k=1}^K (x_k!)^{-1} \theta_k^{x_k}$	$\mathbf{x} \in \mathbb{H}_N^{K-1}$	$\boldsymbol{\theta} \in \Delta^{K-1}$
Beta	$\text{Beta}(x; a, b) \equiv \frac{1}{\mathcal{B}(a,b)} \cdot x^{a-1} (1-x)^{b-1}$	$x \in [0, 1]$	$a > 0, b > 0$
Dirichlet	$\text{Dirichlet}_K(\mathbf{x}; \boldsymbol{\phi}) \equiv \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)} \cdot \prod_{k=1}^K x_k^{\phi_k-1}$	$\mathbf{x} \in \Delta^{K-1}$	$\boldsymbol{\phi} \in \mathbb{R}_{++}^K$

$$\text{Wishart}_M(\mathbf{X}; \mathbf{V}, \nu) \equiv \frac{1}{(2^\nu |\mathbf{V}|)^{M/2} \Gamma_M\left(\frac{\nu}{2}\right)} \cdot \det(\mathbf{X})^{\frac{\nu-M-1}{2}} \exp\left(-\frac{\text{tr}(\mathbf{V}^{-1} \mathbf{X})}{2}\right), \quad (1.18)$$

$$\text{Binomial}_N(x; \theta) \equiv \binom{N}{x} \cdot \theta^x (1-\theta)^{N-x}, \quad (1.19)$$

$$\text{Multinomial}_{K,N}(\mathbf{x}; \boldsymbol{\theta}) \equiv N! \cdot \prod_{k=1}^K (x_k!)^{-1} \theta_k^{x_k}, \quad (1.20)$$

$$\text{Beta}(x; a, b) \equiv \frac{1}{B(a, b)} \cdot x^{a-1} (1-x)^{b-1}, \quad (1.21)$$

$$\text{Dirichlet}_K(\mathbf{x}; \boldsymbol{\phi}) \equiv \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)} \cdot \prod_{k=1}^K x_k^{\phi_{k-1}}. \quad (1.22)$$

The distributions in Table 1.1 are categorized into four groups, which are separated by dashed lines. In each group, an upper distribution family is a special case of a lower distribution family. Note that the following hold:

$$\begin{aligned} \text{Gamma}(x; \alpha, \beta) &= \text{Wishart}_1\left(x; \frac{1}{2\beta}, 2\alpha\right), \\ \text{Binomial}_N(x; \theta) &= \text{Multinomial}_{2,N}\left((x, N-x)^\top; (\theta, 1-\theta)^\top\right), \\ \text{Beta}(x; a, b) &= \text{Dirichlet}_2\left((x, 1-x)^\top; (a, b)^\top\right). \end{aligned}$$

1.2.2 Conjugacy

Let us think about the *function form* of the posterior (1.3):

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \propto p(\mathcal{D}|\mathbf{w})p(\mathbf{w}),$$

which is determined by the function form of the product of the model likelihood $p(\mathcal{D}|\mathbf{w})$ and the prior $p(\mathbf{w})$. Note that we here call the conditional $p(\mathcal{D}|\mathbf{w})$ NOT the *model distribution* but the *model likelihood*, since we are interested in the function form of the posterior, a distribution of the parameter \mathbf{w} .

Conjugacy is defined as the relation between the likelihood $p(\mathcal{D}|\mathbf{w})$ and the prior $p(\mathbf{w})$.

Definition 1.4 (Conjugate prior) A prior $p(\mathbf{w})$ is called *conjugate* with a likelihood $p(\mathcal{D}|\mathbf{w})$, if the posterior $p(\mathbf{w}|\mathcal{D})$ is in the same distribution family as the prior.

1.2.3 Posterior Distribution

Here, we introduce computation of the posterior distribution in simple cases where a conjugate prior exists and is adopted.

Isotropic Gaussian Model

Let us compute the posterior distribution for the isotropic Gaussian model:

$$p(\mathbf{x}|\mathbf{w}) = \text{Gauss}_M(\mathbf{x}; \boldsymbol{\mu}, \sigma^2 \mathbf{I}_M) = \frac{1}{(2\pi\sigma^2)^{M/2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \boldsymbol{\mu}\|^2\right). \quad (1.23)$$

The likelihood for N i.i.d. samples $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ is written as

$$p(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) = \frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2\right)}{(2\pi\sigma^2)^{MN/2}}. \quad (1.24)$$

Gaussian Likelihood As noted in Section 1.2.2, we should see Eq. (1.24), which is the distribution of observed data \mathcal{D} , as a function of the parameter \mathbf{w} . Naturally, the function form depends on which parameters are estimated in the *Bayesian way*. The isotropic Gaussian has two parameters $\mathbf{w} = (\boldsymbol{\mu}, \sigma^2)$, and we first consider the case where the variance parameter σ^2 is known, and the posterior of the mean parameter $\boldsymbol{\mu}$ is estimated, i.e., we set $\mathbf{w} = \boldsymbol{\mu}$. This case contains the case where σ^2 is unknown but point-estimated in the empirical Bayesian procedure or tuned outside the Bayesian framework, e.g., by performing cross-validation (we set $\mathbf{w} = \boldsymbol{\mu}, \kappa = \sigma^2$ in the latter case).

Omitting the constant (with respect to $\boldsymbol{\mu}$), the likelihood (1.24) can be written as

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\mu}) &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|(\mathbf{x}^{(n)} - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\mu})\|^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2 + N\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|^2\right)\right) \\ &\propto \exp\left(-\frac{N}{2\sigma^2} \|\boldsymbol{\mu} - \bar{\mathbf{x}}\|^2\right) \\ &\propto \text{Gauss}_M\left(\boldsymbol{\mu}; \bar{\mathbf{x}}, \frac{\sigma^2}{N} \mathbf{I}_M\right), \end{aligned} \quad (1.25)$$

where $\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)}$ is the *sample mean*. Note that we omitted the factor $\exp(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2)$ as a constant in the fourth equation.

The last equation (1.25) implies that, as a function of the mean parameter μ , the model likelihood $p(\mathcal{D}|\mu)$ has the same form as the isotropic Gaussian with mean \bar{x} and variance $\frac{\sigma^2}{N}$. Eq. (1.25) also implies that the ML estimator for the mean parameter is given by

$$\hat{\mu}^{\text{ML}} = \bar{x}.$$

Thus, we found that the likelihood function for the mean parameter of the isotropic Gaussian is in the Gaussian form. This comes from the following facts:

- The isotropic Gaussian model for a single sample x is in the Gaussian form also as a function of the mean parameter, i.e., $\text{Gauss}_M(x; \mu, \sigma^2 I_M) \propto \text{Gauss}_M(\mu; x, \sigma^2 I_M)$.
- The isotropic Gaussians are *multiplicatively closed*, i.e., the product of isotropic Gaussians with different means is a Gaussian: $p(\mathcal{D}|\mu) \propto \prod_{n=1}^N \text{Gauss}_M(\mu; x^{(n)}, \sigma^2 I_M) \propto \text{Gauss}_M\left(\mu; \bar{x}, \frac{\sigma^2}{N} I_M\right)$.

Since the isotropic Gaussian is multiplicatively closed and the likelihood (1.25) is in the Gaussian form, the isotropic Gaussian prior must be conjugate. Let us choose the isotropic Gaussian prior,

$$p(\mu|\mu_0, \sigma_0^2) = \text{Gauss}_M(\mu; \mu_0, \sigma_0^2 I_M) \propto \exp\left(-\frac{1}{2\sigma_0^2} \|\mu - \mu_0\|^2\right),$$

for hyperparameters $\kappa = (\mu_0, \sigma_0^2)$. Then, the function form of the posterior is given by

$$\begin{aligned} p(\mu|\mathcal{D}, \mu_0, \sigma_0^2) &\propto p(\mathcal{D}|\mu)p(\mu|\mu_0, \sigma_0^2) \\ &\propto \text{Gauss}_M\left(\mu; \bar{x}, \frac{\sigma^2}{N}\right) \text{Gauss}_M(\mu; \mu_0, \sigma_0^2) \\ &\propto \exp\left(-\frac{N}{2\sigma^2} \|\mu - \bar{x}\|^2 - \frac{1}{2\sigma_0^2} \|\mu - \mu_0\|^2\right) \\ &\propto \exp\left(-\frac{N\sigma^{-2} + \sigma_0^{-2}}{2} \left\| \mu - \frac{N\sigma^{-2}\bar{x} + \sigma_0^{-2}\mu_0}{N\sigma^{-2} + \sigma_0^{-2}} \right\|^2\right) \\ &\propto \text{Gauss}_M\left(\mu; \frac{N\sigma^{-2}\bar{x} + \sigma_0^{-2}\mu_0}{N\sigma^{-2} + \sigma_0^{-2}}, \frac{1}{N\sigma^{-2} + \sigma_0^{-2}}\right). \end{aligned}$$

Therefore, the posterior is

$$p(\mu|\mathcal{D}, \mu_0, \sigma_0^2) = \text{Gauss}_M\left(\mu; \frac{N\sigma^{-2}\bar{x} + \sigma_0^{-2}\mu_0}{N\sigma^{-2} + \sigma_0^{-2}}, \frac{1}{N\sigma^{-2} + \sigma_0^{-2}}\right). \quad (1.26)$$

Note that the *equality* holds in Eq. (1.26). We omitted constant factors in the preceding derivation. But once the function form of the posterior is found, the normalization factor is unique. If the function form coincides with one of the well-known distributions (e.g., ones given in Table 1.1), one can find the normalization constant (from the table) without any further computation.

Multiplicative closedness of a function family of the model likelihood is essential in performing Bayesian learning. Such families are called the *exponential family*:

Definition 1.5 (Exponential families) A family of distributions is called the exponential family if it is written as

$$p(\mathbf{x}|\mathbf{w}) = p(\mathbf{t}|\boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^\top \mathbf{t} - A(\boldsymbol{\eta}) + B(\mathbf{t})\right), \quad (1.27)$$

where $\mathbf{t} = \mathbf{t}(\mathbf{x})$ is a function, called *sufficient statistics*, of the random variable \mathbf{x} , and $\boldsymbol{\eta} = \boldsymbol{\eta}(\mathbf{w})$ is a function, called *natural parameters*, of the parameter \mathbf{w} .

The essential property of the exponential family is that the interaction between the random variable and the parameter occurs only in the log linear form, i.e., $\exp(\boldsymbol{\eta}^\top \mathbf{t})$. Note that, although $A(\cdot)$ and $B(\cdot)$ are arbitrary functions, $A(\cdot)$ does not depend on \mathbf{t} , and $B(\cdot)$ does not depend on $\boldsymbol{\eta}$.

Assume that N observed samples $\mathcal{D} = (\mathbf{t}^{(1)}, \dots, \mathbf{t}^{(N)}) = (\mathbf{t}(\mathbf{x}^{(1)}), \dots, \mathbf{t}(\mathbf{x}^{(N)}))$ are drawn from the exponential family distribution (1.27). If we use the exponential family prior $p(\boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^\top \mathbf{t}^{(0)} - A_0(\boldsymbol{\eta}) + B_0(\mathbf{t}^{(0)}))$, then the posterior is given as an exponential family distribution with the same set of natural parameters $\boldsymbol{\eta}$:

$$p(\boldsymbol{\eta}|\mathcal{D}) = \exp\left(\boldsymbol{\eta}^\top \sum_{n=0}^N \mathbf{t}^{(n)} - A'(\boldsymbol{\eta}) + B'(\mathcal{D})\right),$$

where $A'(\boldsymbol{\eta})$ and $B'(\mathcal{D})$ are a function of $\boldsymbol{\eta}$ and a function of \mathcal{D} , respectively. Therefore, the conjugate prior for the exponential family distribution is the exponential family with the same natural parameters $\boldsymbol{\eta}$.

All distributions given in Table 1.1 are exponential families. For example, the sufficient statistics and the natural parameters for the univariate Gaussian are given by $\boldsymbol{\eta} = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})^\top$ and $\mathbf{t} = (x, x^2)^\top$, respectively. The mixture model (1.11) is a common *nonexponential* family distribution.

Gamma Likelihood Next we consider the posterior distribution of the variance parameter σ^2 with the mean parameter regarded as a constant, i.e., $w = \sigma^2$.

Omitting the constants (with respect to σ^2) of the model likelihood (1.24), we have

$$p(\mathcal{D}|\sigma^2) \propto (\sigma^2)^{-MN/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2\right).$$

If we see the likelihood as a function of the *inverse* of σ^2 , we find that it is proportional to the *Gamma distribution*:

$$\begin{aligned} p(\mathcal{D}|\sigma^{-2}) &\propto (\sigma^{-2})^{MN/2} \exp\left(-\left(\frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2\right)\sigma^{-2}\right) \\ &\propto \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2} + 1, \frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2\right). \end{aligned} \quad (1.28)$$

Since the mode of the Gamma distribution is known as $\text{argmax}_x \text{Gamma}(x; \alpha, \beta) = \frac{\alpha-1}{\beta}$, Eq. (1.28) implies that the ML estimator for the variance parameter is given by

$$\widehat{\sigma}^2 \text{ML} = \frac{1}{\widehat{\sigma}^{-2} \text{ML}} = \frac{\frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2}{\frac{MN}{2} + 1 - 1} = \frac{1}{MN} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2.$$

Now we found that the model likelihood of the isotropic Gaussian is in the Gamma form as a function of the inverse variance σ^{-2} . Since the Gamma distribution is in the exponential family and multiplicatively closed, the Gamma prior is conjugate.

If we use the Gamma prior

$$p(\sigma^{-2}|\alpha_0, \beta_0) = \text{Gamma}(\sigma^{-2}; \alpha_0, \beta_0) \propto (\sigma^{-2})^{\alpha_0-1} \exp(-\beta_0 \sigma^{-2})$$

with hyperparameters $\kappa = (\alpha_0, \beta_0)$, the posterior can be written as

$$\begin{aligned} p(\sigma^{-2}|\mathcal{D}, \alpha_0, \beta_0) &\propto p(\mathcal{D}|\sigma^{-2})p(\sigma^{-2}|\alpha_0, \beta_0) \\ &\propto \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2} + 1, \frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2\right) \text{Gamma}(\sigma^{-2}; \alpha_0, \beta_0) \\ &\propto (\sigma^{-2})^{MN/2+\alpha_0-1} \exp\left(-\left(\frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2 + \beta_0\right)\sigma^{-2}\right), \end{aligned}$$

and therefore

$$p(\sigma^{-2}|\mathcal{D}, \alpha_0, \beta_0) = \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2} + \alpha_0, \frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2 + \beta_0\right). \quad (1.29)$$

Isotropic Gauss-Gamma Likelihood Finally, we consider the general case where both the mean and variance parameters are unknown, i.e., $\mathbf{w} = (\boldsymbol{\mu}, \sigma^{-2})$. The likelihood is written as

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\mu}, \sigma^{-2}) &\propto (\sigma^{-2})^{MN/2} \exp\left(-\left(\frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2\right) \sigma^{-2}\right) \\ &= (\sigma^{-2})^{MN/2} \exp\left(-\left(\frac{N\|\boldsymbol{\mu} - \bar{\mathbf{x}}\|^2}{2} + \frac{\sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2}{2}\right) \sigma^{-2}\right) \\ &\propto \text{GaussGamma}_M\left(\boldsymbol{\mu}, \sigma^{-2} \middle| \bar{\mathbf{x}}, N\mathbf{I}_M, \frac{M(N-1)}{2} + 1, \frac{\sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2}{2}\right), \end{aligned}$$

where

$$\begin{aligned} \text{GaussGamma}_M(\mathbf{x}, \tau|\boldsymbol{\mu}, \lambda\mathbf{I}_M, \alpha, \beta) &\equiv \text{Gauss}_M(\mathbf{x}|\boldsymbol{\mu}, (\tau\lambda)^{-1}\mathbf{I}_M) \cdot \text{Gamma}(\tau|\alpha, \beta) \\ &= \frac{\exp\left(-\frac{\tau\lambda}{2}\|\mathbf{x} - \boldsymbol{\mu}\|^2\right)}{(2\pi(\tau\lambda)^{-1})^{M/2}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\beta\tau) \\ &= \frac{\beta^\alpha}{(2\pi/\lambda)^{M/2}\Gamma(\alpha)} \tau^{\alpha+\frac{M}{2}-1} \exp\left(-\left(\frac{\lambda\|\mathbf{x} - \boldsymbol{\mu}\|^2}{2} + \beta\right)\tau\right) \end{aligned}$$

is the *isotropic Gauss-Gamma distribution* on the random variable $\mathbf{x} \in \mathbb{R}^M$, $\tau > 0$ with parameters $\boldsymbol{\mu} \in \mathbb{R}^M$, $\lambda > 0$, $\alpha > 0$, $\beta > 0$.

Note that, although the isotropic Gauss-Gamma distribution is the product of an isotropic Gaussian distribution and a Gamma distribution, the random variables \mathbf{x} and τ are not independent of each other. This is because the isotropic Gauss-Gamma distribution is a *hierarchical model* $p(\mathbf{x}|\tau)p(\tau)$, where the variance parameter $\sigma^2 = (\tau\lambda)^{-1}$ for the isotropic Gaussian depends on the random variable τ of the Gamma distribution.

Since the isotropic Gauss-Gamma distribution is multiplicatively closed, it is a conjugate prior. Choosing the isotropic Gauss-Gamma prior

$$\begin{aligned} p(\boldsymbol{\mu}, \sigma^{-2}|\boldsymbol{\mu}_0, \lambda_0, \alpha_0, \beta_0) &= \text{GaussGamma}_M(\boldsymbol{\mu}, \sigma^{-2}|\boldsymbol{\mu}_0, \lambda_0\mathbf{I}_M, \alpha_0, \beta) \\ &\propto (\sigma^{-2})^{\alpha_0+\frac{M}{2}-1} \exp\left(-\left(\frac{\lambda_0\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2}{2} + \beta_0\right)\sigma^{-2}\right) \end{aligned}$$

with hyperparameters $\boldsymbol{\kappa} = (\boldsymbol{\mu}_0, \lambda_0, \alpha_0, \beta_0)$, the posterior is given by

$$\begin{aligned} p(\boldsymbol{\mu}, \sigma^{-2}|\mathcal{D}, \boldsymbol{\kappa}) &\propto p(\mathcal{D}|\boldsymbol{\mu}, \sigma^{-2})p(\boldsymbol{\mu}, \sigma^{-2}|\boldsymbol{\kappa}) \\ &\propto \text{GaussGamma}_M\left(\boldsymbol{\mu}, \sigma^{-2} \middle| \bar{\mathbf{x}}, N\mathbf{I}_M, \frac{M(N-1)}{2} + 1, \frac{\sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2}{2}\right) \\ &\quad \cdot \text{GaussGamma}_M(\boldsymbol{\mu}, \sigma^{-2}|\boldsymbol{\mu}_0, \lambda_0\mathbf{I}_M, \alpha_0, \beta) \end{aligned}$$

$$\begin{aligned}
& \propto (\sigma^{-2})^{MN/2} \exp \left(- \left(\frac{N\|\boldsymbol{\mu} - \bar{\mathbf{x}}\|^2}{2} + \frac{\sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2}{2} \right) \sigma^{-2} \right) \\
& \quad \cdot (\sigma^{-2})^{\alpha_0 + \frac{M}{2} - 1} \exp \left(- \left(\frac{\lambda_0 \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2}{2} + \beta_0 \right) \sigma^{-2} \right) \\
& \propto (\sigma^{-2})^{M(N+1)/2 + \alpha_0 - 1} \\
& \quad \cdot \exp \left(- \left(\frac{N\|\boldsymbol{\mu} - \bar{\mathbf{x}}\|^2 + \lambda_0 \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2}{2} + \frac{\sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2}{2} + \beta_0 \right) \sigma^{-2} \right) \\
& \propto (\sigma^{-2})^{\widehat{\alpha} + \frac{M}{2} - 1} \exp \left(- \left(\frac{\widehat{\lambda} \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2}{2} + \widehat{\beta} \right) \sigma^{-2} \right),
\end{aligned}$$

where

$$\widehat{\boldsymbol{\mu}} = \frac{N\bar{\mathbf{x}} + \lambda_0 \boldsymbol{\mu}_0}{N + \lambda_0},$$

$$\widehat{\lambda} = N + \lambda_0,$$

$$\widehat{\alpha} = \frac{MN}{2} + \alpha_0,$$

$$\widehat{\beta} = \frac{\sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2}{2} + \frac{N\lambda_0 \|\bar{\mathbf{x}} - \boldsymbol{\mu}_0\|^2}{2(N + \lambda_0)} + \beta_0.$$

Thus, the posterior is obtained as

$$p(\boldsymbol{\mu}, \sigma^{-2} | \mathcal{D}, \boldsymbol{\kappa}) = \text{GaussGamma}_M(\boldsymbol{\mu}, \sigma^{-2} | \widehat{\boldsymbol{\mu}}, \widehat{\lambda} \mathbf{I}_M, \widehat{\alpha}, \widehat{\beta}). \quad (1.30)$$

Although the Gauss-Gamma distribution seems a bit more complicated than the ones in Table 1.1, its moments are known. Therefore, Bayesian learning with a conjugate prior can be analytically performed also when both parameters $\boldsymbol{w} = (\boldsymbol{\mu}, \sigma^{-2})$ are estimated.

Gaussian Model

Bayesian learning can be performed for a general Gaussian model in a similar fashion to the isotropic case. Consider the M -dimensional Gaussian distribution,

$$p(\mathbf{x} | \boldsymbol{w}) = \text{Gauss}_M(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \cdot \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \quad (1.31)$$

with mean and covariance parameters $\boldsymbol{w} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The likelihood for N i.i.d. samples $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ is written as

$$p(\mathcal{D} | \boldsymbol{w}) = \prod_{n=1}^N p(\mathbf{x}^{(n)} | \boldsymbol{w}) = \frac{\exp \left(-\frac{1}{2} \sum_{n=1}^N (\mathbf{x}^{(n)} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(n)} - \boldsymbol{\mu}) \right)}{(2\pi)^{NM/2} \det(\boldsymbol{\Sigma})^{N/2}}. \quad (1.32)$$

Gaussian Likelihood Let us first compute the posterior distribution on the mean parameter μ , with the covariance parameter regarded as a known constant. In this case, the likelihood can be written as

$$\begin{aligned}
p(\mathcal{D}|\mu) &\propto \exp\left(-\frac{1}{2}\sum_{n=1}^N(\mathbf{x}^{(n)} - \mu)^\top \Sigma^{-1}(\mathbf{x}^{(n)} - \mu)\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{n=1}^N((\mathbf{x}^{(n)} - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mu))^\top \cdot \Sigma^{-1}((\mathbf{x}^{(n)} - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mu))\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{n=1}^N(\mathbf{x}^{(n)} - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x}^{(n)} - \bar{\mathbf{x}}) + N(\bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mu)\right)\right) \\
&\propto \exp\left(-\frac{N}{2}(\mu - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mu - \bar{\mathbf{x}})\right) \\
&\propto \text{Gauss}_M\left(\mu; \bar{\mathbf{x}}, \frac{1}{N}\Sigma\right).
\end{aligned} \tag{1.33}$$

Therefore, with the conjugate Gaussian prior

$$p(\mu|\mu_0, \Sigma_0) = \text{Gauss}_M(\mu; \mu_0, \Sigma_0) \propto \exp\left(-\frac{1}{2}(\mu - \mu_0)^\top \Sigma_0^{-1}(\mu - \mu_0)\right),$$

with hyperparameters $\kappa = (\mu_0, \Sigma_0)$, the posterior is written as

$$\begin{aligned}
p(\mu|\mathcal{D}, \mu_0, \Sigma_0) &\propto p(\mathcal{D}|\mu)p(\mu|\mu_0, \Sigma_0) \\
&\propto \text{Gauss}_M\left(\mu; \bar{\mathbf{x}}, \frac{1}{N}\Sigma\right) \text{Gauss}_M(\mu; \mu_0, \Sigma_0) \\
&\propto \exp\left(-\frac{N(\mu - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mu - \bar{\mathbf{x}}) + (\mu - \mu_0)^\top \Sigma_0^{-1}(\mu - \mu_0)}{2}\right) \\
&\propto \exp\left(-\frac{(\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu})}{2}\right),
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\mu} &= \left(N\Sigma^{-1} + \Sigma_0^{-1}\right)^{-1} \left(N\Sigma^{-1}\bar{\mathbf{x}} + \Sigma_0^{-1}\mu_0\right), \\
\widehat{\Sigma} &= \left(N\Sigma^{-1} + \Sigma_0^{-1}\right)^{-1}.
\end{aligned}$$

Thus, we have

$$p(\mu|\mathcal{D}, \mu_0, \Sigma_0) = \text{Gauss}_M\left(\mu; \widehat{\mu}, \widehat{\Sigma}\right). \tag{1.34}$$

Wishart Likelihood If we see the mean parameter μ as a given constant, the model likelihood (1.32) can be written as follows, as a function of the covariance parameter Σ :

$$\begin{aligned} p(\mathcal{D}|\Sigma^{-1}) &\propto \det(\Sigma^{-1})^{N/2} \exp\left(-\frac{\sum_{n=1}^N (\mathbf{x}^{(n)} - \mu)^\top \Sigma^{-1} (\mathbf{x}^{(n)} - \mu)}{2}\right) \\ &\propto \det(\Sigma^{-1})^{N/2} \exp\left(-\frac{\text{tr}(\sum_{n=1}^N (\mathbf{x}^{(n)} - \mu)(\mathbf{x}^{(n)} - \mu)^\top \Sigma^{-1})}{2}\right) \\ &\propto \text{Wishart}_M\left(\Sigma^{-1}; \left(\sum_{n=1}^N (\mathbf{x}^{(n)} - \mu)(\mathbf{x}^{(n)} - \mu)^\top\right)^{-1}, M + N + 1\right). \end{aligned}$$

Here, as in the isotropic Gaussian case, we computed the distribution on the *inverse* Σ^{-1} of the covariance parameter. With the *Wishart distribution*

$$\begin{aligned} p(\Sigma^{-1}|\mathbf{V}_0, \nu_0) &= \text{Wishart}_M(\Sigma^{-1}; \mathbf{V}_0, \nu_0) \\ &= \frac{1}{(2^{\nu_0} \det(\mathbf{V}_0))^{M/2} \Gamma_M\left(\frac{\nu_0}{2}\right)} \cdot \det(\Sigma^{-1})^{\frac{\nu_0-M-1}{2}} \exp\left(-\frac{\text{tr}(\mathbf{V}_0^{-1} \Sigma^{-1})}{2}\right) \end{aligned}$$

for hyperparameters $\kappa = (\mathbf{V}_0, \nu_0)$ as a conjugate prior, the posterior is computed as

$$\begin{aligned} p(\Sigma^{-1}|\mathcal{D}, \mathbf{V}_0, \nu_0) &\propto p(\mathcal{D}|\Sigma^{-1})p(\Sigma^{-1}|\mathbf{V}_0, \nu_0) \\ &\propto \text{Wishart}_M\left(\Sigma^{-1}; \left(\sum_{n=1}^N (\mathbf{x}^{(n)} - \mu)(\mathbf{x}^{(n)} - \mu)^\top\right)^{-1}, M + N + 1\right) \\ &\quad \cdot \text{Wishart}_M(\Sigma^{-1}; \mathbf{V}_0, \nu_0) \\ &\propto \det(\Sigma^{-1})^{\frac{N}{2}} \exp\left(-\frac{\text{tr}((\sum_{n=1}^N (\mathbf{x}^{(n)} - \mu)(\mathbf{x}^{(n)} - \mu)^\top) \Sigma^{-1})}{2}\right) \\ &\quad \cdot \det(\Sigma^{-1})^{\frac{\nu_0-M-1}{2}} \exp\left(-\frac{\text{tr}(\mathbf{V}_0^{-1} \Sigma^{-1})}{2}\right) \\ &\propto \det(\Sigma^{-1})^{\frac{\nu_0-M+N-1}{2}} \exp\left(-\frac{\text{tr}((\sum_{n=1}^N (\mathbf{x}^{(n)} - \mu)(\mathbf{x}^{(n)} - \mu)^\top + \mathbf{V}_0^{-1}) \Sigma^{-1})}{2}\right). \end{aligned}$$

Thus we have

$$\begin{aligned} p(\Sigma^{-1}|\mathcal{D}, \mathbf{V}_0, \nu_0) &= \text{Wishart}_M\left(\Sigma^{-1}; \left(\sum_{n=1}^N (\mathbf{x}^{(n)} - \mu)(\mathbf{x}^{(n)} - \mu)^\top + \mathbf{V}_0^{-1}\right)^{-1}, N + \nu_0\right). \quad (1.35) \end{aligned}$$

Note that the Wishart distribution can be seen as a multivariate extension of the Gamma distribution and is reduced to the Gamma distribution for $M = 1$:

$$\text{Wishart}_1(x; V, \nu) = \text{Gamma}(x; \nu/2, 1/(2V)).$$

Gauss-Wishart Likelihood When both parameters $w = (\mu, \Sigma^{-1})$ are unknown, the model likelihood (1.32) is seen as

$$\begin{aligned} p(\mathcal{D}|\mu, \Sigma^{-1}) &\propto \det(\Sigma^{-1})^{N/2} \exp\left(-\frac{\sum_{n=1}^N (\mathbf{x}^{(n)} - \mu)^\top \Sigma^{-1} (\mathbf{x}^{(n)} - \mu)}{2}\right) \\ &\propto \det(\Sigma^{-1})^{N/2} \exp\left(-\frac{\text{tr}(\sum_{n=1}^N (\mathbf{x}^{(n)} - \mu)(\mathbf{x}^{(n)} - \mu)^\top \Sigma^{-1})}{2}\right) \\ &\propto \det(\Sigma^{-1})^{N/2} \exp\left(-\frac{\text{tr}(\sum_{n=1}^N ((\mathbf{x}^{(n)} - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mu))((\mathbf{x}^{(n)} - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mu))^\top \Sigma^{-1})}{2}\right) \\ &\propto \det(\Sigma^{-1})^{N/2} \exp\left(-\frac{\text{tr}(N(\mu - \bar{\mathbf{x}})(\mu - \bar{\mathbf{x}})^\top + \sum_{n=1}^N (\mathbf{x}^{(n)} - \bar{\mathbf{x}})(\mathbf{x}^{(n)} - \bar{\mathbf{x}})^\top \Sigma^{-1})}{2}\right) \\ &\propto \text{GaussWishart}_M(\mu, \Sigma^{-1}; \bar{\mathbf{x}}, N, (\sum_{n=1}^N (\mathbf{x}^{(n)} - \bar{\mathbf{x}})(\mathbf{x}^{(n)} - \bar{\mathbf{x}})^\top)^{-1}, M + N), \end{aligned}$$

where

$$\begin{aligned} \text{GaussWishart}_M(\mathbf{x}, \Lambda | \mu, \lambda, V, \nu) \\ &\equiv \text{Gauss}_M(\mathbf{x} | \mu, (\lambda \Lambda)^{-1}) \text{Wishart}_M(\Lambda | V, \nu) \\ &= \frac{\exp\left(-\frac{\lambda}{2} (\mathbf{x} - \mu)^\top \Lambda (\mathbf{x} - \mu)\right)}{(2\pi)^{M/2} \det(\lambda \Lambda)^{-1/2}} \cdot \frac{\det(\Lambda)^{\frac{\nu-M-1}{2}} \exp\left(-\frac{\text{tr}(V^{-1} \Lambda)}{2}\right)}{(2^\nu \det(V))^{M/2} \Gamma_M\left(\frac{\nu}{2}\right)} \\ &= \frac{\lambda^{M/2}}{(2^{\nu+1} \pi \det(V))^{M/2} \Gamma_M\left(\frac{\nu}{2}\right)} \det(\Lambda)^{\frac{\nu-M}{2}} \exp\left(-\frac{\text{tr}((\lambda(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top + V^{-1}) \Lambda)}{2}\right) \end{aligned}$$

is the *Gauss–Wishart distribution* on the random variables $\mathbf{x} \in \mathbb{R}^M, \Lambda \in \mathbb{S}_{++}^M$ with parameters $\mu \in \mathbb{R}^M, \lambda > 0, V \in \mathbb{S}_{++}^M, \nu > M - 1$.

With the conjugate Gauss–Wishart prior,

$$\begin{aligned} p(\mu, \Sigma^{-1} | \mu_0, \lambda_0, \alpha_0, \beta_0) &= \text{GaussWishart}_M(\mu, \Sigma^{-1} | \mu_0, \lambda_0, V_0, \nu_0) \\ &\propto \det(\Sigma^{-1})^{\frac{\nu-M}{2}} \exp\left(-\frac{\text{tr}((\lambda_0(\mu - \mu_0)(\mu - \mu_0)^\top + V_0^{-1}) \Sigma^{-1})}{2}\right) \end{aligned}$$

with hyperparameters $\kappa = (\mu_0, \lambda_0, V_0, \nu_0)$, the posterior is written as

$$\begin{aligned} p(\mu, \Sigma^{-1} | \mathcal{D}, \kappa) &\propto p(\mathcal{D} | \mu, \Sigma^{-1}) p(\mu, \Sigma^{-1} | \kappa) \\ &\propto \text{GaussWishart}_M(\mu, \Sigma^{-1}; \bar{\mathbf{x}}, N, (\sum_{n=1}^N (\mathbf{x}^{(n)} - \bar{\mathbf{x}})(\mathbf{x}^{(n)} - \bar{\mathbf{x}})^\top)^{-1}, M + N) \\ &\quad \cdot \text{GaussWishart}_M(\mu, \Sigma^{-1} | \mu_0, \lambda_0, V_0, \nu_0) \end{aligned}$$

$$\begin{aligned} &\propto \det(\boldsymbol{\Sigma}^{-1})^{N/2} \exp\left(-\frac{\text{tr}(N(\mu - \bar{x})(\mu - \bar{x})^\top + \sum_{n=1}^N (\mathbf{x}^{(n)} - \bar{x})(\mathbf{x}^{(n)} - \bar{x})^\top) \boldsymbol{\Sigma}^{-1})}{2}\right) \\ &\quad \cdot \det(\boldsymbol{\Sigma}^{-1})^{\frac{\nu_0 - M}{2}} \exp\left(-\frac{\text{tr}((\lambda_0(\mu - \mu_0)(\mu - \mu_0)^\top + V_0^{-1}) \boldsymbol{\Sigma}^{-1})}{2}\right) \\ &\propto \det(\boldsymbol{\Sigma}^{-1})^{\frac{\widehat{\nu} - M}{2}} \exp\left(-\text{tr}\left(\frac{(\widehat{\lambda}(\mu - \widehat{\mu})(\mu - \widehat{\mu})^\top \widehat{V}^{-1}) \boldsymbol{\Sigma}^{-1}}{2}\right)\right), \end{aligned}$$

where

$$\widehat{\mu} = \frac{N\bar{x} + \lambda_0\mu_0}{N + \lambda_0},$$

$$\widehat{\lambda} = N + \lambda_0,$$

$$\widehat{V} = \left(\sum_{n=1}^N (\mathbf{x}^{(n)} - \bar{x})(\mathbf{x}^{(n)} - \bar{x})^\top + \frac{N\lambda_0}{N + \lambda_0} (\bar{x} - \mu_0)(\bar{x} - \mu_0)^\top + V_0^{-1} \right)^{-1},$$

$$\widehat{\nu} = N + \nu_0.$$

Thus, we have the posterior distribution as the Gauss–Wishart distribution:

$$p(\mu, \boldsymbol{\Sigma}^{-1} | \mathcal{D}, \kappa) = \text{GaussWishart}_M(\mu, \boldsymbol{\Sigma}^{-1} | \widehat{\mu}, \widehat{\lambda}, \widehat{V}, \widehat{\nu}). \quad (1.36)$$

Linear Regression Model

Consider the *linear regression model*, where an input variable $\mathbf{x} \in \mathbb{R}^M$ and an output variable $y \in \mathbb{R}$ are assumed to satisfy the following probabilistic relation:

$$y = \mathbf{a}^\top \mathbf{x} + \varepsilon, \quad (1.37)$$

$$p(\varepsilon | \sigma^2) = \text{Gauss}_1(\varepsilon; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right). \quad (1.38)$$

Here \mathbf{a} and σ^2 are called the *regression parameter* and the *noise variance parameter*, respectively. By substituting $\varepsilon = y - \mathbf{a}^\top \mathbf{x}$, which is obtained from Eq. (1.37), into Eq. (1.38), we have

$$p(y | \mathbf{x}, \mathbf{w}) = \text{Gauss}_1(y; \mathbf{a}^\top \mathbf{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(y - \mathbf{a}^\top \mathbf{x})^2}{2\sigma^2}\right).$$

The likelihood function for N observed i.i.d.² samples,

$$\mathcal{D} = (\mathbf{y}, \mathbf{X}),$$

² In the context of regression, i.i.d. usually means that the observation noise $\varepsilon^{(n)} = y^{(n)} - \mathbf{a}^\top \mathbf{x}^{(n)}$ is independent for different samples, i.e., $p(\{\varepsilon^{(n)}\}_{n=1}^N) = \prod_{n=1}^N p(\varepsilon^{(n)})$, and the independence between the input $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$, i.e., $p(\{\mathbf{x}^{(n)}\}_{n=1}^N) = \prod_{n=1}^N p(\mathbf{x}^{(n)})$, is not required.

is given by

$$p(\mathcal{D}|\mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \cdot \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2}{2\sigma^2}\right), \quad (1.39)$$

where we defined

$$\mathbf{y} = (y^{(1)}, \dots, y^{(N)})^\top \in \mathbb{R}^N, \quad \mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})^\top \in \mathbb{R}^{N \times M}.$$

Gaussian Likelihood The computation of the posterior is similar to the isotropic Gaussian case. As in Section 1.2.3, we first consider the case where only the regression parameter \mathbf{a} is estimated, with the noise variance parameter σ^2 regarded as a known constant.

One can guess that the likelihood (1.39) is Gaussian as a function of \mathbf{a} , since it is an exponential of a concave quadratic function. Indeed, by expanding the exponent and completing the square for \mathbf{a} , we obtain

$$\begin{aligned} p(\mathcal{D}|\mathbf{a}) &\propto \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{(\mathbf{a} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{a} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y})}{2\sigma^2}\right) \\ &\propto \text{Gauss}_M(\mathbf{a}; (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}). \end{aligned} \quad (1.40)$$

Eq. (1.40) implies that, when $\mathbf{X}^\top \mathbf{X}$ is *nonsingular* (i.e., its inverse exists), the ML estimator for \mathbf{a} is given by

$$\widehat{\mathbf{a}}^{\text{ML}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (1.41)$$

Therefore, with the conjugate Gaussian prior

$$p(\mathbf{a}|\mathbf{a}_0, \Sigma_0) = \text{Gauss}_M(\mathbf{a}; \mathbf{a}_0, \Sigma_0) \propto \exp\left(-\frac{1}{2}(\mathbf{a} - \mathbf{a}_0)^\top \Sigma_0^{-1}(\mathbf{a} - \mathbf{a}_0)\right)$$

for hyperparameters $\kappa = (\mathbf{a}_0, \Sigma_0)$, the posterior is Gaussian:

$$\begin{aligned} p(\mathbf{a}|\mathcal{D}, \mathbf{a}_0, \Sigma_0) &\propto p(\mathcal{D}|\mathbf{a})p(\mathbf{a}|\mathbf{a}_0, \Sigma_0) \\ &\propto \text{Gauss}_M(\mathbf{a}; \mathbf{a}_0, \frac{1}{N}\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}) \text{Gauss}_M(\mathbf{a}; \mathbf{a}_0, \Sigma_0) \\ &\propto \exp\left(-\frac{\frac{(\mathbf{a} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{a} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y})}{\sigma^2} + (\mathbf{a} - \mathbf{a}_0)^\top \Sigma_0^{-1}(\mathbf{a} - \mathbf{a}_0)}{2}\right) \\ &\propto \exp\left(-\frac{(\mathbf{a} - \widehat{\mathbf{a}})^\top \widehat{\Sigma}_a^{-1}(\mathbf{a} - \widehat{\mathbf{a}})}{2}\right), \end{aligned}$$

where

$$\begin{aligned}\widehat{\boldsymbol{a}} &= \left(\frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1} \left(\frac{\mathbf{X}^\top \mathbf{y}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1} \mathbf{a}_0 \right), \\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}} &= \left(\frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1}.\end{aligned}$$

Thus we have

$$p(\boldsymbol{a}|\mathcal{D}, \mathbf{a}_0, \boldsymbol{\Sigma}_0) = \text{Gauss}_M(\boldsymbol{a}; \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}). \quad (1.42)$$

Gamma Likelihood When only the noise variance parameter σ^2 is unknown, the model likelihood (1.39) is in the Gamma form, as a function of the inverse σ^{-2} :

$$\begin{aligned}p(\mathcal{D}|\sigma^{-2}) &\propto (\sigma^{-2})^{NM/2} \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{a}\|^2}{2}\sigma^{-2}\right) \\ &\propto \text{Gamma}\left(\sigma^{-2}; \frac{NM}{2} + 1, \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{a}\|^2}{2}\right),\end{aligned} \quad (1.43)$$

which implies that the ML estimator is

$$\widehat{\sigma}^2 \text{ML} = \frac{1}{\widehat{\sigma}^{-2} \text{ML}} = \frac{1}{MN} \sum_{n=1}^N \|\mathbf{y} - \mathbf{X}\boldsymbol{a}\|^2.$$

With the conjugate Gamma prior

$$p(\sigma^{-2}|\alpha_0, \beta_0) = \text{Gamma}(\sigma^{-2}; \alpha_0, \beta_0) \propto (\sigma^{-2})^{\alpha_0-1} \exp(-\beta_0 \sigma^{-2})$$

with hyperparameters $\boldsymbol{\kappa} = (\alpha_0, \beta_0)$, the posterior is computed as

$$\begin{aligned}p(\sigma^{-2}|\mathcal{D}, \alpha_0, \beta_0) &\propto p(\mathcal{D}|\sigma^{-2})p(\sigma^{-2}|\alpha_0, \beta_0) \\ &\propto \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2} + 1, \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{a}\|^2\right) \text{Gamma}(\sigma^{-2}; \alpha_0, \beta_0) \\ &\propto (\sigma^{-2})^{MN/2+\alpha_0-1} \exp\left(-\left(\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{a}\|^2 + \beta_0\right)\sigma^{-2}\right).\end{aligned}$$

Therefore,

$$p(\sigma^{-2}|\mathcal{D}, \alpha_0, \beta_0) = \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2} + \alpha_0, \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{a}\|^2 + \beta_0\right). \quad (1.44)$$

Gauss-Gamma Likelihood When we estimate both parameters $\boldsymbol{w} = (\boldsymbol{a}, \sigma^{-2})$, the likelihood (1.39) is written as

$$\begin{aligned}
p(\mathcal{D}|\mathbf{a}, \sigma^{-2}) &\propto (\sigma^{-2})^{NM/2} \exp\left(-\frac{\|\mathbf{y}-X\mathbf{a}\|^2}{2}\sigma^{-2}\right) \\
&\propto (\sigma^{-2})^{NM/2} \exp\left(-\frac{(\mathbf{a}-\widehat{\mathbf{a}}^{\text{ML}})^T X^\top X(\mathbf{a}-\widehat{\mathbf{a}}^{\text{ML}}) + \|\mathbf{y}-X\widehat{\mathbf{a}}^{\text{ML}}\|^2}{2}\sigma^{-2}\right) \\
&\propto \text{GaussGamma}_M\left(\mathbf{a}, \sigma^{-2}; \widehat{\mathbf{a}}^{\text{ML}}, X^\top X, \frac{M(N-1)}{2} + 1, \frac{\|\mathbf{y}-X\widehat{\mathbf{a}}^{\text{ML}}\|^2}{2}\right),
\end{aligned}$$

where $\widehat{\mathbf{a}}^{\text{ML}}$ is the ML estimator, given by Eq. (1.41), for the regression parameter, and

$$\begin{aligned}
&\text{GaussGamma}_M(\mathbf{x}, \tau|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta) \\
&\equiv \text{Gauss}_M(\mathbf{x}|\boldsymbol{\mu}, (\tau\boldsymbol{\Lambda})^{-1}) \cdot \text{Gamma}(\tau|\alpha, \beta) \\
&= \frac{\exp\left(-\frac{\tau}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right)}{(2\pi\tau^{-1})^{M/2} \det(\boldsymbol{\Lambda})^{-1/2}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\beta\tau) \\
&= \frac{\beta^\alpha}{(2\pi)^{M/2} \det(\boldsymbol{\Lambda})^{-1/2} \Gamma(\alpha)} \tau^{\alpha+\frac{M}{2}-1} \exp\left(-\left(\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2} + \beta\right)\tau\right)
\end{aligned}$$

is the (general) Gauss-Gamma distribution on the random variable $\mathbf{x} \in \mathbb{R}^M$, $\tau > 0$ with parameters $\boldsymbol{\mu} \in \mathbb{R}^M$, $\boldsymbol{\Lambda} \in \mathbb{S}_{++}^M$, $\alpha > 0$, $\beta > 0$. With the conjugate Gauss-Gamma prior

$$\begin{aligned}
p(\mathbf{a}, \sigma^{-2}|\boldsymbol{\kappa}) &= \text{GaussGamma}_M(\mathbf{a}, \sigma^{-2}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \alpha_0, \beta_0) \\
&\propto (\sigma^{-2})^{\alpha_0+\frac{M}{2}-1} \exp\left(-\left(\frac{(\mathbf{a}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0(\mathbf{a}-\boldsymbol{\mu}_0)}{2} + \beta_0\right)\sigma^{-2}\right)
\end{aligned}$$

for hyperparameters $\boldsymbol{\kappa} = (\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \alpha_0, \beta_0)$, the posterior is computed as

$$\begin{aligned}
p(\mathbf{a}, \sigma^{-2}|\mathcal{D}, \boldsymbol{\kappa}) &\propto p(\mathcal{D}|\mathbf{a}, \sigma^{-2})p(\mathbf{a}, \sigma^{-2}|\boldsymbol{\kappa}) \\
&\propto \text{GaussGamma}_M\left(\mathbf{a}, \sigma^{-2}; \widehat{\mathbf{a}}^{\text{ML}}, X^\top X, \frac{M(N-1)}{2} + 1, \frac{\|\mathbf{y}-X\widehat{\mathbf{a}}^{\text{ML}}\|^2}{2}\right) \\
&\quad \cdot \text{GaussGamma}_M(\mathbf{a}, \sigma^{-2}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \alpha_0, \beta_0) \\
&\propto (\sigma^{-2})^{NM/2} \exp\left(-\frac{(\mathbf{a}-\widehat{\mathbf{a}}^{\text{ML}})^T X^\top X(\mathbf{a}-\widehat{\mathbf{a}}^{\text{ML}}) + \|\mathbf{y}-X\widehat{\mathbf{a}}^{\text{ML}}\|^2}{2}\sigma^{-2}\right) \\
&\quad \cdot (\sigma^{-2})^{\alpha_0+\frac{M}{2}-1} \exp\left(-\left(\frac{(\mathbf{a}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0(\mathbf{a}-\boldsymbol{\mu}_0)}{2} + \beta_0\right)\sigma^{-2}\right) \\
&\propto (\sigma^{-2})^{\widehat{\alpha}+\frac{M}{2}-1} \exp\left(-\left(\frac{(\mathbf{a}-\widehat{\boldsymbol{\mu}})^\top \widehat{\boldsymbol{\Lambda}}(\mathbf{a}-\widehat{\boldsymbol{\mu}})}{2} + \widehat{\beta}\right)\sigma^{-2}\right),
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\mu}} &= (X^\top X + \boldsymbol{\Lambda}_0)^{-1} (X^\top X \widehat{\mathbf{a}}^{\text{ML}} + \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0), \\
\widehat{\boldsymbol{\Lambda}} &= X^\top X + \boldsymbol{\Lambda}_0, \\
\widehat{\alpha} &= \frac{NM}{2} + \alpha_0, \\
\widehat{\beta} &= \frac{\|\mathbf{y}-X\widehat{\mathbf{a}}^{\text{ML}}\|^2}{2} + \frac{(\widehat{\mathbf{a}}^{\text{ML}} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0 (X^\top X + \boldsymbol{\Lambda}_0)^{-1} X^\top X (\widehat{\mathbf{a}}^{\text{ML}} - \boldsymbol{\mu}_0)}{2} + \beta_0.
\end{aligned}$$

Thus, we obtain

$$p(\boldsymbol{a}, \sigma^{-2} | \mathcal{D}, \boldsymbol{\kappa}) = \text{GaussGamma}_M(\boldsymbol{a}, \sigma^{-2} | \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Lambda}}, \widehat{\alpha}, \widehat{\beta}). \quad (1.45)$$

Multinomial Model

The *multinomial distribution*, which expresses a distribution over the *histograms* of independent events, is another frequently used basic component in Bayesian modeling. For example, it appears in *mixture models* and *latent Dirichlet allocation*.

Assume that exclusive K events occur with the probability

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_K) \in \Delta^{K-1} \equiv \left\{ \boldsymbol{\theta} \in \mathbb{R}^K; 0 \leq \theta_k \leq 1, \sum_{k=1}^K \theta_k = 1 \right\}.$$

Then, the histogram

$$\mathbf{x} = (x_1, \dots, x_K) \in \mathbb{H}_N^{K-1} \equiv \left\{ \mathbf{x} \in \mathbb{I}^K; 0 \leq x_k \leq N; \sum_{k=1}^K x_k = N \right\}$$

of events after N iterations follows the *multinomial distribution*, defined as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \text{Multinomial}_{K,N}(\mathbf{x}; \boldsymbol{\theta}) \equiv N! \cdot \prod_{k=1}^K \frac{\theta_k^{x_k}}{x_k!}. \quad (1.46)$$

$\boldsymbol{\theta}$ is called the *multinomial parameter*.

As seen shortly, calculation of the posterior with its conjugate prior is surprisingly easy.

Dirichlet Likelihood As a function of the multinomial parameter $\mathbf{w} = \boldsymbol{\theta}$, it is easy to find that the likelihood (1.46) is in the form of the *Dirichlet distribution*:

$$p(\mathbf{x}|\boldsymbol{\theta}) \propto \text{Dirichlet}_K(\boldsymbol{\theta}; \mathbf{x} + \mathbf{1}_K),$$

where $\mathbf{1}_K$ is the K -dimensional vector with all elements equal to 1. Since the Dirichlet distribution is an exponential family and hence multiplicatively closed, it is conjugate for the multinomial parameter. With the conjugate Dirichlet prior

$$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \text{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{\phi}) \propto \prod_{k=1}^K \theta_k^{\phi_k - 1}$$

with hyperparameters $\boldsymbol{\kappa} = \boldsymbol{\phi}$, the posterior is computed as

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\phi}) &\propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi}) \\
&\propto \text{Dirichlet}_K(\boldsymbol{\theta}; \mathbf{x} + \mathbf{1}_K) \cdot \text{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{\phi}) \\
&\propto \prod_{k=1}^K \theta_k^{x_k} \cdot \theta_k^{\phi_k - 1} \\
&\propto \prod_{k=1}^K \theta_k^{x_k + \phi_k - 1}.
\end{aligned}$$

Thus we have

$$p(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\phi}) = \text{Dirichlet}_K(\boldsymbol{\theta}; \mathbf{x} + \boldsymbol{\phi}). \quad (1.47)$$

Special Cases For $K = 2$, the multinomial distribution is reduced to the *binomial distribution*:

$$\begin{aligned}
p(x_1|\theta_1) &= \text{Multinomial}_{2,N}\left((x_1, N - x_1)^\top; (\theta_1, 1 - \theta_1)^\top\right) \\
&= \text{Binomial}_N(x_1; \theta_1) \\
&= \binom{N}{x_1} \cdot \theta_1^{x_1} (1 - \theta_1)^{N - x_1}.
\end{aligned}$$

Furthermore, it is reduced to the *Bernoulli distribution* for $K = 2$ and $N = 1$:

$$\begin{aligned}
p(x_1|\theta_1) &= \text{Binomial}_1(x_1; \theta_1) \\
&= \theta_1^{x_1} (1 - \theta_1)^{1 - x_1}.
\end{aligned}$$

Similarly, its conjugate Dirichlet distribution for $K = 2$ is reduced to the *Beta distribution*:

$$\begin{aligned}
p(\theta_1|\phi_1, \phi_2) &= \text{Dirichlet}_2\left((\theta_1, 1 - \theta_1)^\top; (\phi_1, \phi_2)^\top\right) \\
&= \text{Beta}(\theta_1; \phi_1, \phi_2) \\
&= \frac{1}{\mathcal{B}(\phi_1, \phi_2)} \cdot \theta_1^{\phi_1 - 1} (1 - \theta_1)^{\phi_2 - 1},
\end{aligned}$$

where $\mathcal{B}(\phi_1, \phi_2) = \frac{\Gamma(\phi_1)\Gamma(\phi_2)}{\Gamma(\phi_1 + \phi_2)}$ is the *Beta function*. Naturally, the Beta distribution is conjugate to the binomial and the Bernoulli distributions, and the posterior can be computed as easily as for the multinomial case.

With a conjugate prior in the form of a popular distribution, the four quantities introduced in Section 1.1.3, i.e., the marginal likelihood, the posterior mean, the posterior covariance, and the predictive distribution, can be obtained analytically. In the following subsections, we show how they are obtained.

Table 1.2 *First and second moments of common distributions.*

Mean(\mathbf{x}) = $\langle \mathbf{x} \rangle_{p(\mathbf{x}|\mathbf{w})}$, $\text{Var}(x) = \langle (x - \text{Mean}(x))^2 \rangle_{p(\mathbf{x}|\mathbf{w})}$,
Cov(\mathbf{x}) = $\langle (\mathbf{x} - \text{Mean}(\mathbf{x}))(\mathbf{x} - \text{Mean}(\mathbf{x}))^\top \rangle_{p(\mathbf{x}|\mathbf{w})}$, $\Psi(z) \equiv \frac{d}{dz} \log \Gamma(z)$:
Digamma function, and $\Psi_m(z) \equiv \frac{d^m}{dz^m} \Psi(z)$: Polygamma function of order m .

$p(\mathbf{x} \mathbf{w})$	First moment	Second moment
Gauss _{M} ($\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}$)	Mean (\mathbf{x}) = $\boldsymbol{\mu}$	Cov (\mathbf{x}) = $\boldsymbol{\Sigma}$
Gamma($x; \alpha, \beta$)	$\text{Mean}(x) = \frac{\alpha}{\beta}$ $\text{Mean}(\log x) = \Psi(\alpha) - \log \beta$	$\text{Var}(x) = \frac{\alpha}{\beta^2}$ $\text{Var}(\log x) = \Psi_1(\alpha)$
Wishart _{M} ($\mathbf{X}; \mathbf{V}, \nu$)	Mean (\mathbf{X}) = $\nu \mathbf{V}$	$\text{Var}(x_{m,m'}) = \nu(V_{m,m'}^2 + V_{m,m} V_{m',m'})$
Multinomial _{K,N} ($\mathbf{x}; \boldsymbol{\theta}$)	Mean (\mathbf{x}) = $N\boldsymbol{\theta}$	$(\text{Cov}(\mathbf{x}))_{k,k'} = \begin{cases} N\theta_k(1-\theta_k) & (k=k') \\ -N\theta_k\theta_{k'} & (k \neq k') \end{cases}$
Dirichlet _{K} ($\mathbf{x}; \boldsymbol{\phi}$)	$\text{Mean}(\mathbf{x}) = \frac{1}{\sum_{k=1}^K \phi_k} \boldsymbol{\phi}$ $\text{Mean}(\log x_k) = \Psi(\phi_k) - \Psi(\sum_{k'=1}^K \phi_{k'})$	$(\text{Cov}(\mathbf{x}))_{k,k'} = \begin{cases} \frac{\phi_k(\tau-\phi_k)}{\tau^2(\tau+1)} & (k=k') \\ -\frac{\phi_k\phi_{k'}}{\tau^2(\tau+1)} & (k \neq k') \end{cases}$ where $\tau = \sum_{k=1}^K \phi_k$

1.2.4 Posterior Mean and Covariance

As seen in Section 1.2.3, by adopting a conjugate prior having a form of one of the common family distributions, such as the one in Table 1.1, we can have the posterior distribution in the same common family.³ In such cases, we can simply use the known form of moments, which are summarized in Table 1.2. For example, the posterior (1.42) for the regression parameter \mathbf{a} (when the noise variance σ^2 is treated as a known constant) is the following Gaussian distribution:

$$p(\mathbf{a}|\mathcal{D}, \mathbf{a}_0, \boldsymbol{\Sigma}_0) = \text{Gauss}_M(\mathbf{a}; \widehat{\mathbf{a}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{a}}),$$

where $\widehat{\mathbf{a}} = \left(\frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1} \left(\frac{\mathbf{X}^\top \mathbf{y}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1} \mathbf{a}_0 \right)$,

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{a}} = \left(\frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1}.$$

³ If we would say that the prior is in the family that contains all possible distributions, this family would be the conjugate prior for any likelihood function, which is however useless. Usually, the notion of the conjugate prior implicitly requires that moments (at least the normalization constant and the first moment) of any family member can be computed analytically.

Therefore, the posterior mean and the posterior covariance are simply given by

$$\begin{aligned}\langle \mathbf{a} \rangle_{p(\mathbf{a}|\mathcal{D}, \mathbf{a}_0, \Sigma_0)} &= \widehat{\mathbf{a}}, \\ \langle (\mathbf{a} - \langle \mathbf{a} \rangle)(\mathbf{a} - \langle \mathbf{a} \rangle)^\top \rangle_{p(\mathbf{a}|\mathcal{D}, \mathbf{a}_0, \Sigma_0)} &= \widehat{\Sigma}_{\mathbf{a}},\end{aligned}$$

respectively. The posterior (1.29) of the (inverse) variance parameter σ^{-2} of the isotropic Gaussian distribution (when the mean parameter μ is treated as a known constant) is the following Gamma distribution:

$$p(\sigma^{-2}|\mathcal{D}, \alpha_0, \beta_0) = \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2} + \alpha_0, \frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \mu\|^2 + \beta_0\right).$$

Therefore, the posterior mean and the posterior variance are given by

$$\begin{aligned}\langle \sigma^{-2} \rangle_{p(\sigma^{-2}|\mathcal{D}, \alpha_0, \beta_0)} &= \frac{\frac{MN}{2} + \alpha_0}{\frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \mu\|^2 + \beta_0}, \\ \langle (\sigma^{-2} - \langle \sigma^{-2} \rangle)^2 \rangle_{p(\sigma^{-2}|\mathcal{D}, \alpha_0, \beta_0)} &= \frac{\frac{MN}{2} + \alpha_0}{(\frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \mu\|^2 + \beta_0)^2},\end{aligned}$$

respectively.

Also in other cases, the posterior mean and the posterior covariances can be easily computed by using Table 1.2, if the form of the posterior distribution is in the table.

1.2.5 Predictive Distribution

The predictive distribution (1.9) for a new data set \mathcal{D}^{new} can be computed analytically, if the posterior distribution is in the exponential family, and hence multiplicatively closed. In this section, we show two exemplary cases, the linear regression model and the multinomial model.

Linear Regression Model

Consider the linear regression model:

$$p(y|\mathbf{x}, \mathbf{a}) = \text{Gauss}_1(y; \mathbf{a}^\top \mathbf{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(y - \mathbf{a}^\top \mathbf{x})^2}{2\sigma^2}\right), \quad (1.48)$$

where only the regression parameter is unknown, i.e., $w = \mathbf{a} \in \mathbb{R}^M$, and the noise variance parameter σ^2 is treated as a known constant. We choose the zero-mean Gaussian as a conjugate prior:

$$p(\mathbf{a}|\mathbf{C}) = \text{Gauss}_M(\mathbf{a}; \mathbf{0}, \mathbf{C}) = \frac{\exp\left(-\frac{1}{2}\mathbf{a}^\top \mathbf{C}^{-1} \mathbf{a}\right)}{(2\pi)^{M/2} \det(\mathbf{C})^{1/2}}, \quad (1.49)$$

where \mathbf{C} is the prior covariance.

When N i.i.d. samples $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, where

$$\mathbf{y} = (y^{(1)}, \dots, y^{(N)})^\top \in \mathbb{R}^N, \quad \mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})^\top \in \mathbb{R}^{N \times M},$$

are observed, the posterior is given by

$$\begin{aligned} p(\mathbf{a}|\mathbf{y}, \mathbf{X}, \mathbf{C}) &= \text{Gauss}_M(\mathbf{a}; \widehat{\mathbf{a}}, \widehat{\Sigma}_{\mathbf{a}}) \\ &= \frac{1}{(2\pi)^{M/2} \det(\widehat{\Sigma}_{\mathbf{a}})^{1/2}} \cdot \exp\left(-\frac{(\mathbf{a} - \widehat{\mathbf{a}})^\top \widehat{\Sigma}_{\mathbf{a}}^{-1} (\mathbf{a} - \widehat{\mathbf{a}})}{2}\right), \end{aligned} \quad (1.50)$$

where

$$\widehat{\mathbf{a}} = \left(\frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} + \mathbf{C}^{-1} \right)^{-1} \frac{\mathbf{X}^\top \mathbf{y}}{\sigma^2} = \widehat{\Sigma}_{\mathbf{a}} \frac{\mathbf{X}^\top \mathbf{y}}{\sigma^2}, \quad (1.51)$$

$$\widehat{\Sigma}_{\mathbf{a}} = \left(\frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} + \mathbf{C}^{-1} \right)^{-1}. \quad (1.52)$$

This is just a special case of the posterior (1.42) for the linear regression model with the most general Gaussian prior.

Now, let us compute the predictive distribution on the output y^* for a new given input \mathbf{x}^* . As defined in Eq. (1.9), the predictive distribution is the expectation value of the model distribution (1.48) (for a new input–output pair) over the posterior distribution (1.50):

$$\begin{aligned} p(y^*|\mathbf{x}^*, \mathbf{y}, \mathbf{X}, \mathbf{C}) &= \langle p(y^*|\mathbf{x}^*, \mathbf{a}) \rangle_{p(\mathbf{a}|\mathbf{y}, \mathbf{X}, \mathbf{C})} \\ &= \int p(y^*|\mathbf{x}^*, \mathbf{a}) p(\mathbf{a}|\mathbf{y}, \mathbf{X}, \mathbf{C}) d\mathbf{a} \\ &= \int \text{Gauss}_1(y^*; \mathbf{a}^\top \mathbf{x}^*, \sigma^2) \text{Gauss}_M(\mathbf{a}; \widehat{\mathbf{a}}, \widehat{\Sigma}_{\mathbf{a}}) d\mathbf{a} \\ &\propto \int \exp\left(-\frac{(y^* - \mathbf{a}^\top \mathbf{x}^*)^2}{2\sigma^2} - \frac{(\mathbf{a} - \widehat{\mathbf{a}})^\top \widehat{\Sigma}_{\mathbf{a}}^{-1} (\mathbf{a} - \widehat{\mathbf{a}})}{2}\right) d\mathbf{a} \\ &\propto \exp\left(-\frac{y^{*2}}{2\sigma^2}\right) \int \exp\left(-\frac{\mathbf{a}^\top \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} + \frac{\mathbf{x}^* \mathbf{x}^{*\top}}{\sigma^2}\right) \mathbf{a} - 2\mathbf{a}^\top \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} \widehat{\mathbf{a}} + \frac{\mathbf{x}^* y^*}{\sigma^2}\right)}{2}\right) d\mathbf{a} \\ &\propto \exp\left(-\frac{\sigma^{-2} y^{*2} - \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} \widehat{\mathbf{a}} + \frac{\mathbf{x}^* y^*}{\sigma^2}\right)^\top \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} + \frac{\mathbf{x}^* \mathbf{x}^{*\top}}{\sigma^2}\right)^{-1} \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} \widehat{\mathbf{a}} + \frac{\mathbf{x}^* y^*}{\sigma^2}\right)}{2}\right) \\ &\quad \cdot \int \exp\left(-\frac{(\mathbf{a} - \check{\mathbf{a}})^\top \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} + \frac{\mathbf{x}^* \mathbf{x}^{*\top}}{\sigma^2}\right) (\mathbf{a} - \check{\mathbf{a}})}{2}\right) d\mathbf{a}, \end{aligned} \quad (1.53)$$

where

$$\check{\mathbf{a}} = \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} + \frac{\mathbf{x}^* \mathbf{x}^{*\top}}{\sigma^2} \right)^{-1} \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} \widehat{\mathbf{a}} + \frac{\mathbf{x}^* y^*}{\sigma^2} \right).$$

Note that, although the preceding computation is similar to the one for the posterior distribution in Section 1.2.3, any factor that depends on y^* cannot be ignored even if it does not depend on \mathbf{a} , since the goal is to obtain the distribution on y^* .

The integrand in Eq. (1.53) coincides with the main part of

$$\text{Gauss}_M \left(\mathbf{a}; \tilde{\mathbf{a}}, \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} + \frac{\mathbf{x}^* \mathbf{x}^{*\top}}{\sigma^2} \right)^{-1} \right)$$

without the normalization factor. Therefore, the integral is the inverse of the normalization factor, i.e.,

$$\int \exp \left(-\frac{(\mathbf{a} - \tilde{\mathbf{a}})^{\top} \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} + \frac{\mathbf{x}^* \mathbf{x}^{*\top}}{\sigma^2} \right) (\mathbf{a} - \tilde{\mathbf{a}})}{2} \right) d\mathbf{a} = (2\pi)^{M/2} \det \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} + \frac{\mathbf{x}^* \mathbf{x}^{*\top}}{\sigma^2} \right)^{-1/2},$$

which is a constant with respect to y^* . Therefore, by using Eqs. (1.51) and (1.52), we have

$$\begin{aligned} p(y^* | \mathbf{x}^*, \mathbf{y}, \mathbf{X}, \mathbf{C}) &\propto \exp \left(-\frac{\sigma^{-2} y^{*2} - \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} \tilde{\mathbf{a}} + \frac{\mathbf{x}^* y^*}{\sigma^2} \right)^{\top} \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} + \frac{\mathbf{x}^* \mathbf{x}^{*\top}}{\sigma^2} \right)^{-1} \left(\widehat{\Sigma}_{\mathbf{a}}^{-1} \tilde{\mathbf{a}} + \frac{\mathbf{x}^* y^*}{\sigma^2} \right)}{2} \right) \\ &\propto \exp \left(-\frac{y^{*2} - (\mathbf{X}^{\top} \mathbf{y} + \mathbf{x}^* y^*)^{\top} (\mathbf{X}^{\top} \mathbf{X} + \mathbf{x}^* \mathbf{x}^{*\top} + \sigma^2 \mathbf{C}^{-1})^{-1} (\mathbf{X}^{\top} \mathbf{y} + \mathbf{x}^* y^*)}{2\sigma^2} \right) \\ &\propto \exp \left(-\frac{1}{2\sigma^2} \left\{ y^{*2} \left(1 - \mathbf{x}^{*\top} \left(\mathbf{X}^{\top} \mathbf{X} + \mathbf{x}^* \mathbf{x}^{*\top} + \sigma^2 \mathbf{C}^{-1} \right)^{-1} \mathbf{x}^* \right) \right. \right. \\ &\quad \left. \left. - 2y^* \mathbf{x}^{*\top} \left(\mathbf{X}^{\top} \mathbf{X} + \mathbf{x}^* \mathbf{x}^{*\top} + \sigma^2 \mathbf{C}^{-1} \right)^{-1} \mathbf{X}^{\top} \mathbf{y} \right\} \right) \\ &\propto \exp \left(-\frac{1 - \mathbf{x}^{*\top} \left(\mathbf{X}^{\top} \mathbf{X} + \mathbf{x}^* \mathbf{x}^{*\top} + \sigma^2 \mathbf{C}^{-1} \right)^{-1} \mathbf{x}^*}{2\sigma^2} \right. \\ &\quad \left. \cdot \left(y^* - \frac{\mathbf{x}^{*\top} \left(\mathbf{X}^{\top} \mathbf{X} + \mathbf{x}^* \mathbf{x}^{*\top} + \sigma^2 \mathbf{C}^{-1} \right)^{-1} \mathbf{X}^{\top} \mathbf{y}}{1 - \mathbf{x}^{*\top} \left(\mathbf{X}^{\top} \mathbf{X} + \mathbf{x}^* \mathbf{x}^{*\top} + \sigma^2 \mathbf{C}^{-1} \right)^{-1} \mathbf{x}^*} \right)^2 \right) \\ &\propto \exp \left(-\frac{(y^* - \widehat{y})^2}{2\widehat{\sigma}_y^2} \right), \end{aligned}$$

where

$$\begin{aligned} \widehat{y} &= \frac{\mathbf{x}^{*\top} \left(\mathbf{X}^{\top} \mathbf{X} + \mathbf{x}^* \mathbf{x}^{*\top} + \sigma^2 \mathbf{C}^{-1} \right)^{-1} \mathbf{X}^{\top} \mathbf{y}}{1 - \mathbf{x}^{*\top} \left(\mathbf{X}^{\top} \mathbf{X} + \mathbf{x}^* \mathbf{x}^{*\top} + \sigma^2 \mathbf{C}^{-1} \right)^{-1} \mathbf{x}^*}, \\ \widehat{\sigma}_y^2 &= \frac{\sigma^2}{1 - \mathbf{x}^{*\top} \left(\mathbf{X}^{\top} \mathbf{X} + \mathbf{x}^* \mathbf{x}^{*\top} + \sigma^2 \mathbf{C}^{-1} \right)^{-1} \mathbf{x}^*}. \end{aligned}$$

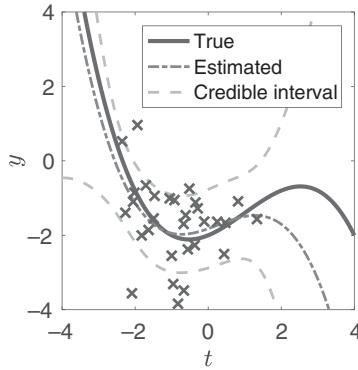


Figure 1.3 Predictive distribution of the linear regression model.

Thus, the predictive distribution has been analytically obtained:

$$p(y^*|\mathbf{x}^*, \mathbf{y}, \mathbf{X}, \mathbf{C}) = \text{Gauss}_1\left(y^*; \hat{y}, \hat{\sigma}_y^2\right). \quad (1.54)$$

Figure 1.3 shows an example of the predictive distribution of the linear regression model. The curve labeled as “True” indicates the mean $y = \mathbf{a}^* \mathbf{x}$ of the true regression model $y = \mathbf{a}^* \mathbf{x} + \varepsilon$, where $\mathbf{a}^* = (-2, 0.4, 0.3, -0.1)^\top$, $\mathbf{x} = (1, t, t^2, t^3)^\top$, and $\varepsilon \sim \text{Gauss}_1(0, 1^2)$. The crosses are $N = 30$ i.i.d. observed samples generated from the true regression model and the input distribution $t \sim \text{Uniform}(-2.4, 1.6)$, where $\text{Uniform}(l, u)$ denotes the uniform distribution on $[l, u]$. The regression model (1.48) with the prior (1.49) for the hyperparameters $\mathbf{C} = 10000 \cdot \mathbf{I}_M, \sigma^2 = 1$ was trained with the observed samples. The curve labeled as “Estimated” and the pair of curves labeled as “Credible interval” show the mean \hat{y} and the *credible interval* $\hat{y} \pm \hat{\sigma}_y$ of the predictive distribution (1.54), respectively.

Reflecting the fact that the samples are observed only in the middle region ($t \in [-2.4, 1.6]$), the credible interval is large in outer regions. The larger interval implies that the “Estimated” function is less reliable, and we see that the gap from the “True” function is indeed large. Since the true function is unknown in practical situations, the variance of the predictive distribution is important information on the reliability of the estimated result.

Multinomial Model

Let us compute the predictive distribution of the multinomial model:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \text{Multinomial}_{K,N}(\mathbf{x}; \boldsymbol{\theta}) \propto \prod_{k=1}^K \frac{\theta_k^{x_k}}{x_k!},$$

$$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \text{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{\phi}) \propto \prod_{k=1}^K \theta_k^{\phi_k - 1},$$

with the observed data $\mathcal{D} = \mathbf{x} = (x_1, \dots, x_K) \in \mathbb{H}_N^{K-1}$ and the unknown parameter $\mathbf{w} = \boldsymbol{\theta} = (\theta_1, \dots, \theta_K) \in \Delta^{K-1}$.

The posterior was derived in Eq. (1.47):

$$p(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\phi}) = \text{Dirichlet}_K(\boldsymbol{\theta}; \mathbf{x} + \boldsymbol{\phi}) \propto \prod_{k=1}^K \theta_k^{x_k + \phi_k - 1}.$$

Therefore, the predictive distribution for a new single sample $\mathbf{x}^* \in \mathbb{H}_1^{K-1}$ is given by

$$\begin{aligned} p(\mathbf{x}^*|\mathbf{x}, \boldsymbol{\phi}) &= \langle p(\mathbf{x}^*|\boldsymbol{\theta}) \rangle_{p(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\phi})} \\ &= \int p(\mathbf{x}^*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\phi}) d\boldsymbol{\theta} \\ &= \int \text{Multinomial}_{K,1}(\mathbf{x}^*; \boldsymbol{\theta}) \text{Dirichlet}_K(\boldsymbol{\theta}; \mathbf{x} + \boldsymbol{\phi}) d\boldsymbol{\theta} \\ &\propto \int \prod_{k=1}^K \theta_k^{x_k^*} \cdot \theta_k^{x_k + \phi_k - 1} d\boldsymbol{\theta} \\ &= \int \prod_{k=1}^K \theta_k^{x_k^* + x_k + \phi_k - 1} d\boldsymbol{\theta}. \end{aligned} \quad (1.55)$$

In the fourth equation, we ignored the factors that depend neither on \mathbf{x}^* nor on $\boldsymbol{\theta}$.

The integrand in Eq. (1.55) is the main part of $\text{Dirichlet}_K(\boldsymbol{\theta}; \mathbf{x}^* + \mathbf{x} + \boldsymbol{\phi})$, and therefore, the integral is equal to the inverse of its normalization factor:

$$\begin{aligned} \int \prod_{k=1}^K \theta_k^{x_k^* + x_k + \phi_k - 1} d\boldsymbol{\theta} &= \frac{\prod_{k=1}^K \Gamma(x_k^* + x_k + \phi_k)}{\Gamma(\sum_{k=1}^K x_k^* + x_k + \phi_k)} \\ &= \frac{\prod_{k=1}^K \Gamma(x_k^* + x_k + \phi_k)}{\Gamma(N + \sum_{k=1}^K \phi_k + 1)}. \end{aligned}$$

Thus, by using the identity $\Gamma(x+1) = x\Gamma(x)$ for the Gamma function, we have

$$\begin{aligned} p(\mathbf{x}^*|\mathbf{x}, \boldsymbol{\phi}) &\propto \prod_{k=1}^K \Gamma(x_k^* + x_k + \phi_k) \\ &\propto \prod_{k=1}^K (x_k + \phi_k)^{x_k^*} \Gamma(x_k + \phi_k) \end{aligned}$$

$$\begin{aligned}
&\propto \prod_{k=1}^K (x_k + \phi_k)^{x_k^*} \\
&\propto \prod_{k=1}^K \left(\frac{x_k + \phi_k}{\sum_{k'=1}^K x_{k'} + \phi_{k'}} \right)^{x_k^*} \\
&= \text{Multinomial}_{K,1}(\mathbf{x}^*; \widehat{\boldsymbol{\theta}}),
\end{aligned} \tag{1.56}$$

where

$$\widehat{\theta}_k = \frac{x_k + \phi_k}{\sum_{k'=1}^K x_{k'} + \phi_{k'}}. \tag{1.57}$$

From Eq. (1.47) and Table 1.2, we can easily see that the predictive mean $\widehat{\boldsymbol{\theta}}$, specified by Eq. (1.57), coincides with the posterior mean, i.e., the Bayesian estimator:

$$\widehat{\boldsymbol{\theta}} = \langle \boldsymbol{\theta} \rangle_{\text{Dirichlet}_K(\boldsymbol{\theta}; \mathbf{x} + \boldsymbol{\phi})}.$$

Therefore, in the multinomial model, the predictive distribution coincides with the model distribution with the Bayesian estimator plugged in.

In the preceding derivation, we performed the integral computation and derived the form of the predictive distribution. However, the necessary information to determine the predictive distribution is the probability table on the events $\mathbf{x}^* \in \mathbb{H}_1^{K-1} = \{\mathbf{e}_k\}_{k=1}^K$, of which the degree of freedom is only K . Therefore, the following simple calculation gives the same result:

$$\begin{aligned}
\text{Prob}(\mathbf{x}^* = \mathbf{e}_k | \mathbf{x}, \boldsymbol{\phi}) &= \langle \text{Multinomial}_{K,1}(\mathbf{e}_k; \boldsymbol{\theta}) \rangle_{\text{Dirichlet}_K(\boldsymbol{\theta}; \mathbf{x} + \boldsymbol{\phi})} \\
&= \langle \theta_k \rangle_{\text{Dirichlet}_K(\boldsymbol{\theta}; \mathbf{x} + \boldsymbol{\phi})} \\
&= \widehat{\theta}_k,
\end{aligned}$$

which specifies the function form of the predictive distribution, given by Eq. (1.56).

1.2.6 Marginal Likelihood

Let us compute the marginal likelihood of the linear regression model, defined by Eqs. (1.48) and (1.49):

$$\begin{aligned}
p(\mathcal{D}|C) &= p(\mathbf{y}|X, \mathbf{C}) \\
&= \langle p(\mathbf{y}|X, \mathbf{a}) \rangle_{p(\mathbf{a}|C)} \\
&= \int p(\mathbf{y}|X, \mathbf{a}) p(\mathbf{a}|C) d\mathbf{a}
\end{aligned}$$

$$\begin{aligned}
&= \int \text{Gauss}_N(\mathbf{y}; \mathbf{X}\mathbf{a}, \sigma^2 \mathbf{I}_N) \text{Gauss}_M(\mathbf{a}; \mathbf{0}, \mathbf{C}) d\mathbf{a} \\
&= \int \frac{\exp\left(-\frac{\|\mathbf{y}-\mathbf{X}\mathbf{a}\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{N/2}} \cdot \frac{\exp\left(-\frac{1}{2}\mathbf{a}^\top \mathbf{C}^{-1}\mathbf{a}\right)}{(2\pi)^{M/2} \det(\mathbf{C})^{1/2}} d\mathbf{a} \\
&= \frac{\exp\left(-\frac{\|\mathbf{y}\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{N/2}(2\pi)^{M/2} \det(\mathbf{C})^{1/2}} \\
&\quad \cdot \int \exp\left(-\frac{-2\mathbf{a}^\top \frac{\mathbf{X}^\top \mathbf{y}}{\sigma^2} + \mathbf{a}^\top \left(\frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} + \mathbf{C}^{-1}\right)\mathbf{a}}{2}\right) d\mathbf{a} \\
&= \frac{\exp\left(-\frac{1}{2}\left(\frac{\|\mathbf{y}\|^2}{\sigma^2} - \widehat{\mathbf{a}}^\top \widehat{\Sigma}_a^{-1} \widehat{\mathbf{a}}\right)\right)}{(2\pi\sigma^2)^{N/2}(2\pi)^{M/2} \det(\mathbf{C})^{1/2}} \\
&\quad \cdot \int \exp\left(-\frac{(\mathbf{a} - \widehat{\mathbf{a}})^\top \widehat{\Sigma}_a^{-1} (\mathbf{a} - \widehat{\mathbf{a}})}{2}\right) d\mathbf{a}, \tag{1.58}
\end{aligned}$$

where $\widehat{\mathbf{a}}$ and $\widehat{\Sigma}_a$ are, respectively, the posterior mean and the posterior covariance, given by Eqs. (1.51) and (1.52).

By using

$$\int \exp\left(-\frac{(\mathbf{a} - \widehat{\mathbf{a}})^\top \widehat{\Sigma}_a^{-1} (\mathbf{a} - \widehat{\mathbf{a}})}{2}\right) d\mathbf{a} = \sqrt{(2\pi)^M \det(\widehat{\Sigma}_a)},$$

and Eq. (1.58), we have

$$\begin{aligned}
p(\mathbf{y}|\mathbf{X}, \mathbf{C}) &= \frac{\exp\left(-\frac{1}{2}\left(\frac{\|\mathbf{y}\|^2}{\sigma^2} - \frac{\mathbf{y}^\top \mathbf{X} \widehat{\Sigma}_a \mathbf{X}^\top \mathbf{y}}{\sigma^4}\right)\right)}{(2\pi\sigma^2)^{N/2}(2\pi)^{M/2} \det(\mathbf{C})^{1/2}} \sqrt{(2\pi)^M \det(\widehat{\Sigma}_a)} \\
&= \frac{\exp\left(-\frac{\|\mathbf{y}\|^2 - \mathbf{y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \sigma^2 \mathbf{C}^{-1})^{-1} \mathbf{X}^\top \mathbf{y}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{N/2} \det(\mathbf{C} \mathbf{X}^\top \mathbf{X} + \sigma^2 \mathbf{I}_M)^{1/2}}, \tag{1.59}
\end{aligned}$$

where we also used Eqs. (1.51) and (1.52).

Eq. (1.59) is an explicit expression of the marginal likelihood as a function of the hyperparameter $\kappa = \mathbf{C}$. Based on it, we perform EBays learning in Section 1.2.7.

1.2.7 Empirical Bayesian Learning

In empirical Bayesian (EBays) learning, the hyperparameter κ is estimated by maximizing the marginal likelihood $p(\mathcal{D}|\kappa)$. The negative logarithm of the marginal likelihood,

$$F^{\text{Bayes}} = -\log p(\mathcal{D}|\kappa), \quad (1.60)$$

is called the *Bayes free energy* or *stochastic complexity*.⁴ Since $\log(\cdot)$ is a monotonic function, maximizing the marginal likelihood is equivalent to minimizing the Bayes free energy.

Eq. (1.59) implies that the Bayes free energy of the linear regression model is given by

$$\begin{aligned} 2F^{\text{Bayes}} &= -2 \log p(\mathbf{y}|\mathbf{X}, \mathbf{C}) \\ &= N \log(2\pi\sigma^2) + \log \det(\mathbf{C}\mathbf{X}^\top \mathbf{X} + \sigma^2 \mathbf{I}_M) \\ &\quad + \frac{\|\mathbf{y}\|^2 - \mathbf{y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \sigma^2 \mathbf{C}^{-1})^{-1} \mathbf{X}^\top \mathbf{y}}{\sigma^2}. \end{aligned} \quad (1.61)$$

Let us restrict the prior covariance to be diagonal:

$$\mathbf{C} = \text{Diag}(c_1^2, \dots, c_M^2) \in \mathbb{D}^M. \quad (1.62)$$

The prior (1.49) with diagonal covariance (1.62) is called the *automatic relevance determination (ARD)* prior, which is known to make the EBayes estimator sparse (Neal, 1996). In the following example, we see this effect by setting the design matrix to identity, $\mathbf{X} = \mathbf{I}_M$, which enables us to derive the EBayes solution analytically.

Under the identity design matrix, the Bayes free energy (1.61) can be decomposed as

$$\begin{aligned} 2F^{\text{Bayes}} &= N \log(2\pi\sigma^2) + \log \det(\mathbf{C} + \sigma^2 \mathbf{I}_M) + \frac{\|\mathbf{y}\|^2 - \mathbf{y}^\top (\mathbf{I}_M + \sigma^2 \mathbf{C}^{-1})^{-1} \mathbf{y}}{\sigma^2} \\ &= N \log(2\pi\sigma^2) + \frac{\|\mathbf{y}\|^2}{\sigma^2} + \sum_{m=1}^M \left(\log(c_m^2 + \sigma^2) - \frac{y_m^2}{\sigma^2(1 + \sigma^2 c_m^{-2})} \right) \\ &= \sum_{m=1}^M 2F_m^* + \text{const.}, \end{aligned} \quad (1.63)$$

where

$$2F_m^* = \log \left(1 + \frac{c_m^2}{\sigma^2} \right) - \frac{y_m^2}{\sigma^2} \left(1 + \frac{\sigma^2}{c_m^2} \right)^{-1}. \quad (1.64)$$

In Eq. (1.63), we omitted the constant factors with respect to the hyperparameter \mathbf{C} . As the remaining terms are decomposed into each component m , we can independently minimize F_m^* with respect to c_m^2 .

⁴ The logarithm of the marginal likelihood $\log p(\mathcal{D}|\kappa)$ is called the *log marginal likelihood* or *evidence*.

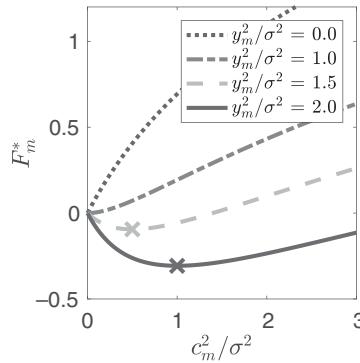


Figure 1.4 The (componentwise) Bayes free energy (1.64) of linear regression model with the ARD prior. The minimizer is shown as a cross if it lies in the positive region of c_m^2 / σ^2 .

The derivative of Eq. (1.64) with respect to c_m^2 is

$$\begin{aligned} 2 \frac{\partial F_m^*}{\partial c_m^2} &= \frac{1}{c_m^2 + \sigma^2} - \frac{y_m^2}{(1 + \sigma^2 c_m^{-2})^2 c_m^4} \\ &= \frac{1}{c_m^2 + \sigma^2} - \frac{y_m^2}{(c_m^2 + \sigma^2)^2} \\ &= \frac{c_m^2 - (y_m^2 - \sigma^2)}{(c_m^2 + \sigma^2)^2}. \end{aligned} \quad (1.65)$$

Eq. (1.65) implies that F_m^* is monotonically increasing over all domain $c_m^2 > 0$ when $y_m^2 \leq \sigma^2$, and has the unique minimizer in the region $c_m^2 > 0$ when $y_m^2 > \sigma^2$. Specifically, the minimizer is given by

$$\widehat{c}_m^2 = \begin{cases} y_m^2 - \sigma^2 & \text{if } y_m^2 > \sigma^2, \\ +0 & \text{otherwise.} \end{cases} \quad (1.66)$$

Figure 1.4 shows the (componentwise) Bayes free energy (1.64) for different observations, $y_m^2 = 0, \sigma^2, 1.5\sigma^2, 2\sigma^2$. The minimizer is in the positive region of c_m^2 if and only if $y_m^2 > \sigma^2$.

If the EBayes estimator is given by $\widehat{c}_m^2 \rightarrow +0$, it means that the *prior* distribution for the m th component a_m of the regression parameter is the *Dirac delta function* located at the origin.⁵ This formally means that we *a priori*

⁵ When $y_m^2 \leq \sigma^2$, the Bayes free energy (1.64) decreases as c_m^2 approaches to 0. However, the domain of c_m^2 is restricted to be positive, and therefore, $\widehat{c}_m^2 = 0$ is not the solution. We express this solution as $\widehat{c}_m^2 \rightarrow +0$.

knew that $a_m = 0$, i.e., we choose a model that does not contain the m th component.

By substituting Eq. (1.66) into the Bayes posterior mean (1.51), we obtain the EBayes estimator:

$$\begin{aligned}\widehat{a}_m^{\text{EBayes}} &= \widehat{c}_m^2 \left(\widehat{c}_m^2 + \sigma^2 \right)^{-1} y_m \\ &= \begin{cases} \left(1 - \frac{\sigma^2}{y_m^2} \right) y_m & \text{if } y_m^2 > \sigma^2, \\ 0 & \text{otherwise.} \end{cases} \quad (1.67)\end{aligned}$$

The form of the estimator (1.67) is called the *James–Stein (JS) estimator* having interesting properties including the *domination* over the ML estimator (Stein, 1956; James and Stein, 1961; Efron and Morris, 1973) (see Appendix A).

Note that the assumption that $\mathbf{X} = \mathbf{I}_M$ is not practical. For a general design matrix \mathbf{X} , the Bayes free energy is not decomposable into each component. Consequently, the prior variances $\{c_m^2\}_{m=1}^M$ that minimize the Bayes free energy (1.61) interact with each other. Therefore, the preceding simple mechanism is not applied. However, it is empirically observed that many prior variances tend to go to $\widehat{c}_m^2 \rightarrow +0$, so that the EBayes estimator $\widehat{\boldsymbol{a}}^{\text{EBayes}}$ is sparse.

2

Variational Bayesian Learning

In Chapter 1, we saw examples where the model likelihood has a conjugate prior, with which Bayesian learning can be performed analytically. However, many practical models do not have conjugate priors. Even in such cases, the notion of conjugacy is still useful. Specifically, we can make use of the *conditional conjugacy*, which comes from the fact that many practical models are built by combining basic distributions. In this chapter, we introduce *variational Bayesian (VB) learning*, which makes use of the conditional conjugacy, and approximates the Bayes posterior by solving a constrained minimization problem.

2.1 Framework

VB learning is derived by casting Bayesian learning as an optimization problem with respect to the posterior distribution (Hinton and van Camp, 1993; MacKay, 1995; Opper and Winther, 1996; Attias, 1999; Jordan et al., 1999; Jaakkola and Jordan, 2000; Ghahramani and Beal, 2001; Bishop, 2006; Wainwright and Jordan, 2008).

2.1.1 Free Energy Minimization

Let $r(\mathbf{w})$, or r for short, be an arbitrary distribution, which we call a *trial distribution*, on the parameter \mathbf{w} , and consider the *Kullback–Leibler (KL) divergence* from the trial distribution $r(\mathbf{w})$ to the Bayes posterior $p(\mathbf{w}|\mathcal{D})$:

$$\text{KL}(r(\mathbf{w})\|p(\mathbf{w}|\mathcal{D})) = \int r(\mathbf{w}) \log \frac{r(\mathbf{w})}{p(\mathbf{w}|\mathcal{D})} d\mathbf{w} = \left\langle \log \frac{r(\mathbf{w})}{p(\mathbf{w}|\mathcal{D})} \right\rangle_{r(\mathbf{w})}. \quad (2.1)$$

Since the KL divergence is equal to zero if and only if the two distributions coincide with each other, the minimizer of Eq. (2.1) is the Bayes posterior, i.e.,

$$p(\mathbf{w}|\mathcal{D}) = \operatorname{argmin}_r \text{KL}(r(\mathbf{w})\|p(\mathbf{w}|\mathcal{D})). \quad (2.2)$$

The problem (2.2) is equivalent to the following problem:

$$p(\mathbf{w}|\mathcal{D}) = \operatorname{argmin}_r F(r), \quad (2.3)$$

where the functional of r ,

$$F(r) = \int r(\mathbf{w}) \log \frac{r(\mathbf{w})}{p(\mathbf{w}, \mathcal{D})} d\mathbf{w} = \left\langle \log \frac{r(\mathbf{w})}{p(\mathbf{w}, \mathcal{D})} \right\rangle_{r(\mathbf{w})} \quad (2.4)$$

$$= \text{KL}(r(\mathbf{w})\|p(\mathbf{w}|\mathcal{D})) - \log p(\mathcal{D}), \quad (2.5)$$

is called the *free energy*. Intuitively, we replaced the posterior distribution $p(\mathbf{w}|\mathcal{D})$ in the KL divergence (2.1) with its unnormalized version—the joint distribution $p(\mathcal{D}, \mathbf{w}) = p(\mathbf{w}|\mathcal{D})p(\mathcal{D})$ —in the free energy (2.4). The equivalence holds because the normalization factor $p(\mathcal{D})$ does not depend on \mathbf{w} , and therefore $\langle \log p(\mathcal{D}) \rangle_{r(\mathbf{w})} = \log p(\mathcal{D})$ does not depend on r . Note that the free energy (2.4) is a generalization of the Bayes free energy, defined by Eq. (1.60): The free energy (2.4) is a functional of an arbitrary distribution r , and equal to the Bayes free energy (1.60) for the Bayes posterior $r(\mathbf{w}) = p(\mathbf{w}|\mathcal{D})$. Since the KL divergence is nonnegative, Eq. (2.5) implies that the free energy $F(r)$ is an upper-bound of the Bayes free energy $-\log p(\mathcal{D})$ for any distribution r . Since the log marginal likelihood $\log p(\mathcal{D})$ is called the evidence, $-F(r)$ is also called the *evidence lower-bound (ELBO)*.

As mentioned in Section 1.1.2, the joint distribution is easy to compute in general. However, the minimization problem in Eq. (2.3) can still be computationally intractable, because the objective functional (2.4) involves the *expectation* over the distribution $r(\mathbf{w})$. Actually, it can be hard to even evaluate the objective functional for most of the possible distributions. To make the evaluation of the objective functional tractable *for optimal $r(\mathbf{w})$* , we restrict the search space to \mathcal{G} . Namely, we solve the following problem:

$$\min_r F(r) \quad \text{s.t.} \quad r \in \mathcal{G}, \quad (2.6)$$

where s.t. is an abbreviation for “subject to.”

We can choose a tractable distribution class directly for \mathcal{G} , e.g., Gaussian, such that the expectation for evaluating the free energy is tractable for any $r \in \mathcal{G}$. However, in many practical models, a weaker constraint restricts the *optimal distribution* to be in a tractable class, thanks to *conditional conjugacy*.

2.1.2 Conditional Conjugacy

Let us consider a few examples where the model likelihood has no conjugate prior. The likelihood of the matrix factorization model (which will be discussed in detail in Section 3.1) is given by

$$p(\mathbf{V}|\mathbf{A}, \mathbf{B}) = \frac{\exp\left(-\frac{1}{2\sigma^2} \|\mathbf{V} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2\right)}{(2\pi\sigma^2)^{LM/2}}, \quad (2.7)$$

where $\mathbf{V} \in \mathbb{R}^{L \times M}$ is an observed random variable, and $\mathbf{A} \in \mathbb{R}^{M \times H}$ and $\mathbf{B} \in \mathbb{R}^{L \times H}$ are the parameters to be estimated. Although $\sigma^2 \in \mathbb{R}_{++}$ can also be unknown, let us treat it as a hyperparameter, i.e., a constant when computing the posterior distribution.

If we see Eq. (2.7) as a function of the parameters $\mathbf{w} = (\mathbf{A}, \mathbf{B})$, its function form is the exponential of a polynomial including a fourth-order term $\|\mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2 = \text{tr}(\mathbf{B}\mathbf{A}^\top \mathbf{A}\mathbf{B}^\top)$. Therefore, no conjugate prior exists for this likelihood with respect to the parameters $\mathbf{w} = (\mathbf{A}, \mathbf{B})$.¹

The next example is a mixture of Gaussians (which will be discussed in detail in Section 4.1.1):

$$p(\mathcal{D}, \mathcal{H}|\mathbf{w}) = \prod_{n=1}^N \prod_{k=1}^K \left\{ \alpha_k \frac{\exp\left(-\frac{\|\mathbf{x}^{(n)} - \boldsymbol{\mu}_k\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{M/2}} \right\}^{z_k^{(n)}}, \quad (2.8)$$

where $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^N$ are observed data, $\mathcal{H} = \{\mathbf{z}^{(n)}\}_{n=1}^N$ are hidden variables, and $\mathbf{w} = (\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)$ are parameters. For simplicity, we here assume that all Gaussian components have the same variance σ^2 , which is treated as a hyperparameter, i.e., we compute the joint posterior distribution of the hidden variables $\{\mathbf{z}^{(n)}\}_{n=1}^N$ and the parameters \mathbf{w} , regarding the hyperparameter σ^2 as a constant.

If we see Eq. (2.8) as a function of $(\{\mathbf{z}^{(n)}\}_{n=1}^N, \boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)$, no conjugate prior exists. More specifically, it has a factor $\prod_{n=1}^N \prod_{k=1}^K \alpha_k^{z_k^{(n)}}$, and we cannot compute

$$\sum_{z_k^{(n)} \in \{e_k\}_{k=1}^K} \int \prod_{n=1}^N \prod_{k=1}^K \alpha_k^{z_k^{(n)}} d\alpha_k$$

analytically for general N , which is required when evaluating moments.

¹ Here, “no conjugate prior” means that there is no *useful* and *nonconditional* conjugate prior, such that the posterior is in the same distribution family with computable moments. We might say that the exponential function of fourth-order polynomials is conjugate to the likelihood (2.7), since the posterior is within the same family. However, this statement is useless in practice because we cannot compute moments of the distribution analytically.

The same difficulty happens in the latent Dirichlet allocation model (which will be discussed in detail in Section 4.2.4). The likelihood is written as

$$p(\mathcal{D}, \mathcal{H}|\mathbf{w}) = \prod_{m=1}^M \prod_{n=1}^{N^{(m)}} \prod_{h=1}^H \left\{ \Theta_{m,h} \prod_{l=1}^L B_{l,h}^{w_l^{(n,m)}} \right\}^{z_h^{(n,m)}}, \quad (2.9)$$

where $\mathcal{D} = \{\{\mathbf{w}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M$ are observed data, $\mathcal{H} = \{\{z^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M$ are hidden variables, and $\mathbf{w} = (\boldsymbol{\Theta}, \mathbf{B})$ are parameters to be estimated. Computing the sum (over the hidden variables \mathcal{H}) of the integral (over the parameters \mathbf{w}) is intractable for practical problem sizes.

Readers might find that Eqs. (2.7), (2.8), and (2.9) are not much more complicated than the conjugate cases: Eq. (2.7) is similar to the Gaussian form, and Eqs. (2.8) and (2.9) are in the form of the multinomial or Dirichlet distribution, where we have unknowns both in the base and in the exponent. Indeed, they are in a known form if we regard a part of unknowns as fixed constants.

The likelihood (2.7) of the matrix factorization model is in the Gaussian form of \mathbf{A} if we see \mathbf{B} as a constant, or vice versa. The likelihood (2.8) of a mixture of Gaussians is in the multinomial form of the hidden variables $\mathcal{H} = \{z^{(n)}\}_{n=1}^N$ if we see the parameters $\mathbf{w} = (\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)$ as constants, and it is the (independent) product of the Dirichlet form of $\boldsymbol{\alpha}$ and the Gaussian form of $\{\boldsymbol{\mu}_k\}_{k=1}^K$ if we see the hidden variables $\mathcal{H} = \{z^{(n)}\}_{n=1}^N$ as constants. Similarly, the likelihood (2.9) of the latent Dirichlet allocation model is in the multinomial form of the hidden variables $\mathcal{H} = \{\{z^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M$ if we see the parameters $\mathbf{w} = (\boldsymbol{\Theta}, \mathbf{B})$ as constants, and it is the product of the Dirichlet form of the row vectors $\{\boldsymbol{\theta}_m\}_{m=1}^M$ of $\boldsymbol{\Theta}$ and the Dirichlet form of the column vectors $\{\boldsymbol{\beta}_h\}_{h=1}^H$ of \mathbf{B} if we see the hidden variables $\mathcal{H} = \{\{z^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M$ as constants.

Since the likelihoods in the Gaussian, multinomial, and Dirichlet forms have conjugate priors, the aforementioned properties can be described with the notion of *conditional conjugacy*, which is defined as follows:

Definition 2.1 (Conditionally conjugate prior) Let us divide the unknown parameters \mathbf{w} (or more generally all unknown variables including hidden variables) into two parts $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2)$. If the posterior of \mathbf{w}_1 ,

$$p(\mathbf{w}_1|\mathbf{w}_2, \mathcal{D}) \propto p(\mathcal{D}|\mathbf{w}_1, \mathbf{w}_2)p(\mathbf{w}_1), \quad (2.10)$$

is in the same distribution family as the prior $p(\mathbf{w}_1)$ (where \mathbf{w}_2 is regarded as a given constant or condition), the prior $p(\mathbf{w}_1)$ is called a *conditionally conjugate prior* of the model likelihood $p(\mathcal{D}|\mathbf{w})$ with respect to the parameter \mathbf{w}_1 , given the fixed parameter \mathbf{w}_2 .

2.1.3 Constraint Design

Once conditional conjugacy for all unknowns is found, designing tractable VB learning is straightforward.

Let us divide the unknown parameters \mathbf{w} into S groups, i.e., $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_S)$, such that, for each $s = 1, \dots, S$, the model likelihood $p(\mathcal{D}|\mathbf{w}) = p(\mathcal{D}|\mathbf{w}_s, \{\mathbf{w}_{s'}\}_{s' \neq s})$ has a conditionally conjugate prior $p(\mathbf{w}_s)$ with respect to \mathbf{w}_s , given $\{\mathbf{w}_{s'}\}_{s' \neq s}$ as fixed constants. Then, if we use the prior

$$p(\mathbf{w}) = \prod_{s=1}^S p(\mathbf{w}_s), \quad (2.11)$$

the posterior distribution

$$p(\mathbf{w}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{w})p(\mathbf{w})$$

is, as a function of \mathbf{w}_s , in the same distribution family as the prior $p(\mathbf{w}_s)$. Therefore, moments of the posterior distribution are tractable, if the other parameters $\{\mathbf{w}_{s'}\}_{s' \neq s}$ are given.

To make use of this property, we impose on the approximate posterior the independence constraint between the parameter groups,

$$r(\mathbf{w}) = \prod_{s=1}^S r_s(\mathbf{w}_s), \quad (2.12)$$

which allows us to compute moments with respect to \mathbf{w}_s independently from the other parameters $\{\mathbf{w}_{s'}\}_{s' \neq s}$. In VB learning, we solve the minimization problem (2.6) under the constraint (2.12). This makes the expectation computation, which is required in evaluating the free energy (2.4), tractable (on any stationary points for r). Namely, we define the *VB posterior* as

$$\widehat{r} = \underset{r}{\operatorname{argmin}} F(r) \quad \text{s.t.} \quad r(\mathbf{w}) = \prod_{s=1}^S r_s(\mathbf{w}_s). \quad (2.13)$$

Note that it is not guaranteed that the free energy $F(r) = \left\langle \log \frac{r(\mathbf{w})}{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})} \right\rangle_{r(\mathbf{w})}$ is tractable for any r satisfying the constraint (2.12). However, the constraint allows us to optimize each factor $\{r_s\}_{s=1}^S$ separately. To optimize each factor, we rely on *calculus of variations*, which will be explained in Section 2.1.4. By applying calculus of variations, the free energy is expressed as an explicit function with a *finite* number of unknown parameters.

2.1.4 Calculus of Variations

Calculus of variations is a method, developed in physics, to derive conditions that any optimal function minimizing a (smooth) functional should satisfy (Courant and Hilbert, 1953). Specifically, it gives (infinitely many) stationary conditions of the functional with respect to the variable.

The change of the functional $F(r)$ with respect to an *infinitesimal change* of the variable r (which is a function of \mathbf{w}) is called a *variation* and written as δI . For r to be a stationary point of the functional, the variation must be equal to zero for all possible values of \mathbf{w} . Since the free energy (2.4) does not depend on the derivatives of $r(\mathbf{w})$, the variation δI is simply the derivative with respect to r . Therefore, the stationary conditions are given by

$$\delta I = \frac{\partial F}{\partial r} = 0, \quad \forall \mathbf{w} \in \mathcal{W}, \quad (2.14)$$

which is a special case of the *Euler–Lagrange equation*. If we see the function $r(\mathbf{w})$ as a (possibly) infinite-dimensional vector with the parameter value \mathbf{w} as its index, the variation $\delta I = \delta I(\mathbf{w})$ can be interpreted as the gradient of the functional $F(r)$ in the $|\mathcal{W}|$ -dimensional space. As the stationary conditions in a finite-dimensional space require that all entries of the gradient equal to zero, the optimal function $r(\mathbf{w})$ should satisfy Eq. (2.14) for any parameter values $\mathbf{w} \in \mathcal{W}$.

In Section 2.1.5, we see that, by applying the stationary conditions (2.14) to the free energy minimization problem (2.13) with the independence constraint taken into account, we can find that each factor $r_s(\mathbf{w}_s)$ of the approximate posterior is in the same distribution family as the corresponding prior $p_s(\mathbf{w}_s)$, thanks to the conditional conjugacy.

2.1.5 Variational Bayesian Learning

Let us solve the problem (2.13) to get the VB posterior

$$\hat{r} = \underset{r}{\operatorname{argmin}} F(r) \quad \text{s.t.} \quad r(\mathbf{w}) = \prod_{s=1}^S r_s(\mathbf{w}_s).$$

We use the decomposable conditionally conjugate prior (2.11):

$$p(\mathbf{w}) = \prod_{s=1}^S p(\mathbf{w}_s),$$

which means that, for each $s = 1, \dots, S$, the posterior $p(\mathbf{w}_s | \{\mathbf{w}_{s'}\}_{s' \neq s}, \mathcal{D})$ for \mathbf{w}_s is in the same form as the corresponding prior $p(\mathbf{w}_s)$, given $\{\mathbf{w}_{s'}\}_{s' \neq s}$ as fixed constants.

Now we apply the calculus of variations, and compute the stationary conditions (2.14). The free energy can be written as

$$F(r) = \int \left(\prod_{s=1}^S r_s(\mathbf{w}_s) \right) \left(\log \frac{\prod_{s=1}^S r_s(\mathbf{w}_s)}{p(\mathcal{D}|\mathbf{w}) \prod_{s=1}^S p(\mathbf{w}_s)} \right) d\mathbf{w}. \quad (2.15)$$

Taking the derivative of Eq. (2.15) with respect to $r_s(\mathbf{w}_s)$ for any $s = 1, \dots, S$ and $\mathbf{w}_s \in \mathcal{W}$, we obtain the following stationary conditions:

$$\begin{aligned} 0 &= \frac{\partial F}{\partial r_s} = \int \left(\prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'}) \right) \left(\log \frac{\prod_{s'=1}^S r_{s'}(\mathbf{w}_{s'})}{p(\mathcal{D}|\mathbf{w}) \prod_{s'=1}^S p(\mathbf{w}_{s'})} + 1 \right) d\mathbf{w} \\ &= \left\langle \log \frac{\prod_{s'=1}^S r_{s'}(\mathbf{w}_{s'})}{p(\mathcal{D}|\mathbf{w}) \prod_{s'=1}^S p(\mathbf{w}_{s'})} \right\rangle_{\prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'})} + 1 \\ &= \left\langle \log \frac{\prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'})}{p(\mathcal{D}|\mathbf{w}) \prod_{s' \neq s} p(\mathbf{w}_{s'})} \right\rangle_{\prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'})} + \log \frac{r_s(\mathbf{w}_s)}{p(\mathbf{w}_s)} + 1 \\ &= \left\langle \log \frac{1}{p(\mathcal{D}|\mathbf{w})} \right\rangle_{\prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'})} + \log \frac{r_s(\mathbf{w}_s)}{p(\mathbf{w}_s)} + \text{const.} \end{aligned} \quad (2.16)$$

Note the following on Eq. (2.16):

- The right-hand side is a function of \mathbf{w}_s ($\mathbf{w}_{s'}$ for $s' \neq s$ are integrated out).
- For each s , Eq. (2.16) must hold for any possible value of \mathbf{w}_s , which can fully specify the function form of the posterior $r_s(\mathbf{w}_s)$.
- To make Eq. (2.16) satisfied for any \mathbf{w}_s , it is necessary that

$$-\langle \log p(\mathcal{D}|\mathbf{w}) \rangle_{\prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'})} + \log \frac{r_s(\mathbf{w}_s)}{p(\mathbf{w}_s)}$$

is a constant.

The last note leads to the following relation:

$$r_s(\mathbf{w}_s) \propto p(\mathbf{w}_s) \exp \langle \log p(\mathcal{D}|\mathbf{w}) \rangle_{\prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'})}. \quad (2.17)$$

As a function of \mathbf{w}_s , Eq. (2.17) can be written as

$$\begin{aligned} r_s(\mathbf{w}_s) &\propto \exp \langle \log p(\mathcal{D}|\mathbf{w}) p(\mathbf{w}_s) \rangle_{\prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'})} \\ &\propto \exp \langle \log p(\mathbf{w}_s | \{\mathbf{w}_{s'}\}_{s' \neq s}, \mathcal{D}) \rangle_{\prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'})} \\ &\propto \exp \int \log p(\mathbf{w}_s | \{\mathbf{w}_{s'}\}_{s' \neq s}, \mathcal{D}) \prod_{s' \neq s} r_{s'}(\mathbf{w}_{s'}) d\mathbf{w}_{s'}. \end{aligned} \quad (2.18)$$

Due to the conditional conjugacy, $p(\mathbf{w}_s | \{\mathbf{w}_{s'}\}_{s' \neq s}, \mathcal{D})$ is in the same form as the prior $p(\mathbf{w}_s)$. As the intergral operator $g(x) = \int f(x; \alpha) d\alpha$ can be interpreted as an infinite number of additions of parametric functions $f(x; \alpha)$ over all possible values of α , the operator $h(x) = \exp \int \log f(x; \alpha) d\alpha$ corresponds to an infinite

number of multiplications of $f(x; \alpha)$ over all possible values of α . Therefore, Eq. (2.18) implies that the VB posterior $r_s(\mathbf{w}_s)$ is in the same form as the prior $p(\mathbf{w}_s)$, if the distribution family is multiplicatively closed.

Assume that the prior $p(\mathbf{w}_s)$ for each group of parameters is in a multiplicatively closed distribution family. Then, we may express the corresponding VB posterior $r_s(\mathbf{w}_s)$ in a parametric form, of which the parameters are called *variational parameters*, without any loss of accuracy or optimality. The last question is whether we can compute the expectation value of the log-likelihood $\log p(\mathcal{D}|\mathbf{w})$ for each factor $r_s(\mathbf{w}_s)$ of the approximate posterior. In many cases, this expectation can be computed analytically, which allows us to express the stationary conditions (2.17) as a finite number of equations in explicit forms of the variational parameters.

Typically, the obtained stationary conditions are used to update the variational parameters in an iterative algorithm, which gives a local minimizer \widehat{r} of the free energy (2.4). We call the minimizer \widehat{r} the *VB posterior*, and its mean

$$\widehat{\mathbf{w}} = \langle \mathbf{w} \rangle_{\widehat{r}(\mathbf{w})} \quad (2.19)$$

the *VB estimator*.

The computation of predictive distribution

$$p(\mathcal{D}^{\text{new}}|\mathcal{D}) = \langle p(\mathcal{D}^{\text{new}}|\mathbf{w}) \rangle_{\widehat{r}(\mathbf{w})}$$

can be hard even after finding the VB posterior $\widehat{r}(\mathbf{w})$. This is natural because we need approximation for the function form of the likelihood $p(\mathcal{D}|\mathbf{w})$, and now we need to compute the integral with the integrand involving the same function form. In many practical cases, the *plug-in predictive distribution* $p(\mathcal{D}^{\text{new}}|\widehat{\mathbf{w}})$, i.e., the model distribution with the VB estimator plugged in, is substituted for the predictive distribution.

2.1.6 Empirical Variational Bayesian Learning

When the model involves hyperparameters κ in the likelihood and/or in the prior, the joint distribution is dependent on κ , i.e.,

$$p(\mathcal{D}, \mathbf{w}|\kappa) = p(\mathbf{w}|\kappa)p(\mathcal{D}|\mathbf{w}, \kappa),$$

and so is the free energy:

$$\begin{aligned} F(r, \kappa) &= \int r(\mathbf{w}) \log \frac{r(\mathbf{w})}{p(\mathcal{D}, \mathbf{w}|\kappa)} d\mathbf{w} \\ &= \left\langle \log \frac{r(\mathbf{w})}{p(\mathbf{w}, \mathcal{D}|\kappa)} \right\rangle_{r(\mathbf{w})} \end{aligned} \quad (2.20)$$

$$= \text{KL}(r(\mathbf{w})||p(\mathbf{w}|\mathcal{D}, \kappa)) - \log p(\mathcal{D}|\kappa). \quad (2.21)$$

Similarly to the empirical Bayesian learning, the hyperparameters can be estimated from observation by minimizing the free energy simultaneously with respect to r and κ :

$$\widehat{(r, \kappa)} = \operatorname{argmin}_{r, \kappa} F(r, \kappa).$$

This approach is called the *empirical VB (EVB) learning*.

EVB learning amounts to minimizing the sum of the KL divergence to the Bayes posterior and the marginal likelihood (see Eq. (2.21)). Conceptually, minimizing any weighted sum of those two terms is reasonable to find the VB posterior and the hyperparameters at the same time. But only the unweighted sum makes the objective tractable—under this choice, the objective is written with the joint distribution as in Eq. (2.20), while any other choice requires explicitly accessing the Bayes posterior and the marginal likelihood separately.

2.1.7 Techniques for Nonconjugate Models

In Sections 2.1.2 through 2.1.5, we saw how to design tractable VB learning by making use of the conditional conjugacy. However, there are also many cases where a reasonable model does not have a conditionally conjugate prior. A frequent and important example is the case where the likelihood involves the *sigmoid function*,

$$\sigma(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}}, \quad (2.22)$$

or a function with a similar shape, e.g., the *error function*, the hyperbolic tangent, and the *rectified linear unit (ReLU)*. We face such cases, for example, in solving *classification* problems and in adopting *neural network* structure with a nonlinear *activation function*.

To maintain the tractability in such cases, we need to explicitly restrict the function form of the approximate posterior $r(\mathbf{w}; \widehat{\lambda})$, and optimize its variational parameters $\widehat{\lambda}$ by free energy minimization. Namely, we solve the VB learning problem (2.6) with the search space \mathcal{G} set to the function space of a simple distribution family, e.g., the Gaussian distribution $r(\mathbf{w}; \widehat{\lambda}) = \text{Gauss}_D(\mathbf{w}; \widehat{\mathbf{w}}, \widehat{\Sigma})$ parameterized with the variational parameters $\widehat{\lambda} = (\widehat{\mathbf{w}}, \widehat{\Sigma})$ consisting of the mean and the covariance parameters. Then, the VB learning problem (2.6) is reduced to the following unconstrained minimization problem,

$$\min_{\widehat{\lambda}} F(\widehat{\lambda}), \quad (2.23)$$

of the free energy

$$F(\widehat{\lambda}) = \int r(\mathbf{w}; \widehat{\lambda}) \log \frac{r(\mathbf{w}; \widehat{\lambda})}{p(\mathbf{w}, \mathcal{D})} d\mathbf{w} = \left\langle \log \frac{r(\mathbf{w}; \widehat{\lambda})}{p(\mathbf{w}, \mathcal{D})} \right\rangle_{r(\mathbf{w}; \widehat{\lambda})}, \quad (2.24)$$

which is a function of the variational parameters $\widehat{\lambda}$.

It is often the case that the free energy (2.24) is still intractable in computing the expectation value of the log joint probability, $\log p(\mathbf{w}, \mathcal{D}) = \log p(\mathcal{D}|\mathbf{w})p(\mathbf{w})$, over the approximate posterior $r(\mathbf{w}; \widehat{\lambda})$ (because of the intractable function form of the likelihood $p(\mathcal{D}|\mathbf{w})$ or the prior $p(\mathbf{w})$). In this section, we introduce a few techniques developed for coping with such intractable functions.

Local Variational Approximation

The first method is to bound the joint distribution $p(\mathbf{w}, \mathcal{D})$ with a simple function, of which the expectation value over the approximate distribution $r(\mathbf{w}; \widehat{\lambda})$ is tractable.

As seen in Section 2.1.1, the free energy (2.24) is an upper-bound of the Bayes free energy, $F^{\text{Bayes}} \equiv -\log p(\mathcal{D})$, for any $\widehat{\lambda}$. Consider further upper-bounding the free energy as

$$\bar{F}(\widehat{\lambda}) \leq \underline{F}(\widehat{\lambda}, \xi) \equiv \int r(\mathbf{w}; \widehat{\lambda}) \log \frac{r(\mathbf{w}; \widehat{\lambda})}{\underline{p}(\mathbf{w}; \xi)} d\mathbf{w} \quad (2.25)$$

by replacing the joint distribution $p(\mathbf{w}, \mathcal{D})$ with its parametric lower-bound $\underline{p}(\mathbf{w}; \xi)$ such that

$$0 \leq \underline{p}(\mathbf{w}; \xi) \leq p(\mathbf{w}, \mathcal{D}) \quad (2.26)$$

for any $\mathbf{w} \in \mathcal{W}$ and $\xi \in \Xi$. Here, we introduced another set of variational parameters ξ with its domain Ξ . Let us choose a lower-bound $\underline{p}(\mathbf{w}; \xi)$ such that its function form with respect to \mathbf{w} is the same as the approximate posterior $r(\mathbf{w}; \widehat{\lambda})$. More specifically, we assume that, for any given ξ , there exists $\widehat{\lambda}$ such that

$$\underline{p}(\mathbf{w}; \xi) \propto r(\mathbf{w}; \widehat{\lambda}) \quad (2.27)$$

as a function of \mathbf{w} .² Since the direct minimization of $\bar{F}(\widehat{\lambda})$ is intractable, we instead minimize its upper-bound $\underline{F}(\widehat{\lambda}, \xi)$ jointly over $\widehat{\lambda}$ and ξ . Namely, we solve the problem

$$\min_{\widehat{\lambda}, \xi} \underline{F}(\widehat{\lambda}, \xi), \quad (2.28)$$

² The parameterization, i.e., the function form with respect to the variational parameters, can be different between $\underline{p}(\mathbf{w}; \xi)$ and $r(\mathbf{w}; \widehat{\lambda})$.

to find the approximate posterior $r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})$ such that $\overline{F}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\xi}) (\geq F(\widehat{\boldsymbol{\lambda}}))$ is closest to the Bayes free energy F^{Bayes} (when $\boldsymbol{\xi}$ is also optimized).

Let

$$q(\mathbf{w}; \boldsymbol{\xi}) = \frac{\underline{p}(\mathbf{w}; \boldsymbol{\xi})}{\underline{Z}(\boldsymbol{\xi})} \quad (2.29)$$

be the distribution created by normalizing the lower-bound with its normalization factor

$$\underline{Z}(\boldsymbol{\xi}) = \int \underline{p}(\mathbf{w}; \boldsymbol{\xi}) d\mathbf{w}. \quad (2.30)$$

Note that the normalization factor (2.30) is trivially a lower-bound of the marginal likelihood, i.e.,

$$\underline{Z}(\boldsymbol{\xi}) \leq \int p(\mathbf{w}, \mathcal{D}) d\mathbf{w} = p(\mathcal{D}),$$

and is tractable because of the assumption (2.27) that \underline{p} is in the same simple function form as r .

With Eq. (2.29), the upper-bound (2.25) is expressed as

$$\begin{aligned} \overline{F}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\xi}) &= \int r(\mathbf{w}; \widehat{\boldsymbol{\lambda}}) \log \frac{r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})}{q(\mathbf{w}; \boldsymbol{\xi})} d\mathbf{w} - \log \underline{Z}(\boldsymbol{\xi}) \\ &= \text{KL}\left(r(\mathbf{w}; \widehat{\boldsymbol{\lambda}}) \| q(\mathbf{w}; \boldsymbol{\xi})\right) - \log \underline{Z}(\boldsymbol{\xi}), \end{aligned} \quad (2.31)$$

which implies that the optimal $\widehat{\boldsymbol{\lambda}}$ is attained when

$$r(\mathbf{w}; \widehat{\boldsymbol{\lambda}}) = q(\mathbf{w}; \boldsymbol{\xi}) \quad (2.32)$$

for any $\boldsymbol{\xi} \in \Xi$ (the assumption (2.27) guarantees the attainability). Thus, by putting this back into Eq. (2.31), the problem (2.28) is reduced to

$$\max_{\boldsymbol{\xi}} \underline{Z}(\boldsymbol{\xi}), \quad (2.33)$$

which amounts to maximizing the lower-bound (2.30) of the marginal likelihood $p(\mathcal{D})$. Once the maximizer $\widehat{\boldsymbol{\xi}}$ is obtained, Eq. (2.32) gives the optimal approximate posterior.

Such an approximation scheme for nonconjugate models is called *local variational approximation* or *direct site bounding* (Jaakkola and Jordan, 2000; Girolami, 2001; Bishop, 2006; Seeger, 2008, 2009), which will be discussed further with concrete examples in Chapter 5. Existing nonconjugate models applied with the local variational approximation form the bound in Eq. (2.26) based on the convexity of a function. In such a case, the gap between $\overline{F}(\boldsymbol{\xi})$ and F turns out to be the expected Bregman divergence associated with the convex function (see Section 5.3.1). A similar approach can be

applied to *expectation propagation*, another approximation method introduced in Section 2.2.3. There, by upper-bounding the joint probability $p(\mathbf{w}, \mathcal{D})$, we minimize an upper-bound of $\text{KL}\left(p(\mathbf{w}|\mathcal{D})\|r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})\right)$ (see Section 2.2.3).

Black Box Variational Inference

As the available data size increases, and the benefit of using big data has been proven, for example, by the breakthrough in deep learning (Krizhevsky et al., 2012), scalable training algorithms have been intensively developed, to enable big data analysis on billions of data samples. The *stochastic gradient descent* (Robbins and Monro, 1951; Spall, 2003), where a noisy gradient of the objective function is cheaply computed from a subset of the whole data in each iteration, has become popular, and has been adopted for VB learning (Hoffman et al., 2013; Khan et al., 2016).

The *black-box variational inference* was proposed as a general method to compute a noisy gradient of the free energy in nonconjugate models (Ranganath et al., 2013; Wingate and Weber, 2013; Kingma and Welling, 2014). As a function of the variational parameters $\widehat{\boldsymbol{\lambda}}$, the gradient of the free energy (2.24) can be evaluated by

$$\begin{aligned}\frac{\partial F}{\partial \widehat{\boldsymbol{\lambda}}} &= \frac{\partial}{\partial \widehat{\boldsymbol{\lambda}}} \int r(\mathbf{w}; \widehat{\boldsymbol{\lambda}}) \log \frac{r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})}{p(\mathcal{D}, \mathbf{w})} d\mathbf{w} \\ &= \int \frac{\partial r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})}{\partial \widehat{\boldsymbol{\lambda}}} \log \frac{r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})}{p(\mathcal{D}, \mathbf{w})} d\mathbf{w} + \int r(\mathbf{w}; \widehat{\boldsymbol{\lambda}}) \frac{\partial}{\partial \widehat{\boldsymbol{\lambda}}} (\log r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})) d\mathbf{w} \\ &= \int r(\mathbf{w}; \widehat{\boldsymbol{\lambda}}) \frac{\partial \log r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})}{\partial \widehat{\boldsymbol{\lambda}}} \log \frac{r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})}{p(\mathcal{D}, \mathbf{w})} d\mathbf{w} + \frac{\partial}{\partial \widehat{\boldsymbol{\lambda}}} \int r(\mathbf{w}; \widehat{\boldsymbol{\lambda}}) d\mathbf{w} \\ &= \left\langle \frac{\partial \log r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})}{\partial \widehat{\boldsymbol{\lambda}}} \log \frac{r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})}{p(\mathcal{D}, \mathbf{w})} \right\rangle_{r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})}. \end{aligned} \quad (2.34)$$

Assume that we restrict the approximate posterior $r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})$ to be in a simple distribution family, from which samples can be easily drawn, and its *score function*, the gradient of the log probability, is easily computed, e.g., an analytic form is available. Then, Eq. (2.34) can be easily computed by drawing samples from $r(\mathbf{w}; \widehat{\boldsymbol{\lambda}})$, and computing the sample average. With some variance reduction techniques, the stochastic gradient with the black box gradient estimator (2.34) has shown to be useful for VB learning in general nonconjugate models. A notable advantage is that it does not require any model specific analysis to implement the gradient estimation, since Eq. (2.34) can be evaluated as long as the log joint probability $p(\mathcal{D}, \mathbf{w}) = p(\mathcal{D}|\mathbf{w})p(\mathbf{w})$ of the model can be evaluated for drawn samples of \mathbf{w} .

2.2 Other Approximation Methods

There are several other methods for approximate Bayesian learning, which are briefly introduced in this section.

2.2.1 Laplace Approximation

In the *Laplace approximation*, the posterior is approximated by a Gaussian:

$$r(\mathbf{w}) = \text{Gauss}_D(\mathbf{w}; \widehat{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}}).$$

VB learning finds the variational parameters $\widehat{\lambda} = (\widehat{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}})$ by minimizing the free energy (2.4), i.e., solving the problem (2.6) with the search space \mathcal{G} restricted to the Gaussian distributions. Instead, the Laplace approximation estimates the mean and the covariance by

$$\widehat{\mathbf{w}}^{\text{LA}} (= \widehat{\mathbf{w}}^{\text{MAP}}) = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}), \quad (2.35)$$

$$\widehat{\boldsymbol{\Sigma}}^{\text{LA}} = \widehat{\mathbf{F}}^{-1}, \quad (2.36)$$

where the entries of $\widehat{\mathbf{F}} \in \mathbb{S}_{++}^D$ are given by

$$\widehat{F}_{i,j} = -\left. \frac{\partial^2 \log p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{\partial w_i \partial w_j} \right|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{LA}}}.$$
(2.37)

Namely, the Laplace approximation first finds the MAP estimator for the mean, and then computes Eq.(2.37) at $\mathbf{w} = \widehat{\mathbf{w}}^{\text{LA}}$ to estimate the inverse covariance, which corresponds to the second-order *Taylor approximation* to $\log p(\mathcal{D}|\mathbf{w})p(\mathbf{w})$. Note that, for the flat prior $p(\mathbf{w}) \propto 1$, Eq. (2.37) is reduced to the *Fisher information*:

$$F_{i,j} = \left\langle \frac{\partial \log p(\mathcal{D}|\mathbf{w})}{\partial w_i} \frac{\partial \log p(\mathcal{D}|\mathbf{w})}{\partial w_j} \right\rangle_{p(\mathcal{D}|\mathbf{w})} = - \left\langle \frac{\partial^2 \log p(\mathcal{D}|\mathbf{w})}{\partial w_i \partial w_j} \right\rangle_{p(\mathcal{D}|\mathbf{w})}.$$

In general, the Laplace approximation is computationally less demanding than VB learning, since no integral computation is involved, and the inverse covariance estimation (2.36) is performed only once after the MAP mean estimator (2.35) is found.

2.2.2 Partially Bayesian Learning

Partially Bayesian (PB) learning is MAP learning after some of the unknown parameters are integrated out. This approach can be described in the free energy minimization framework (2.6) with a stronger constraint than VB learning.

Let us split the unknown parameters \mathbf{w} into two parts $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2)$, and assume that we integrate \mathbf{w}_1 out and point-estimate \mathbf{w}_2 . Integrating \mathbf{w}_1 out means that we consider the exact posterior on \mathbf{w}_1 , and MAP estimating \mathbf{w}_2 means that we approximate the posterior \mathbf{w}_2 with the delta function. Namely, PB learning solves the following problem:

$$\min_r F(r) \quad \text{s.t.} \quad r(\mathbf{w}) = r_1(\mathbf{w}_1) \cdot \delta(\mathbf{w}_2; \widehat{\mathbf{w}}_2), \quad (2.38)$$

where the free energy $F(r)$ is defined by Eq. (2.4), and $\delta(\mathbf{w}; \widehat{\mathbf{w}})$ is the *Dirac delta function* located at $\widehat{\mathbf{w}}$.

Using the constraint in Eq. (2.38), under which the variables to be optimized are r_1 and $\widehat{\mathbf{w}}_2$, we can express the free energy as

$$\begin{aligned} F(r_1, \widehat{\mathbf{w}}_2) &= \left\langle \log \frac{r_1(\mathbf{w}_1) \cdot \delta(\mathbf{w}_2; \widehat{\mathbf{w}}_2)}{p(\mathcal{D}|\mathbf{w}_1, \mathbf{w}_2)p(\mathbf{w}_1)p(\mathbf{w}_2)} \right\rangle_{r_1(\mathbf{w}_1), \delta(\mathbf{w}_2; \widehat{\mathbf{w}}_2)} \\ &= \left\langle \log \frac{r_1(\mathbf{w}_1)}{p(\mathcal{D}|\mathbf{w}_1, \widehat{\mathbf{w}}_2)p(\mathbf{w}_1)p(\widehat{\mathbf{w}}_2)} \right\rangle_{r_1(\mathbf{w}_1)} + \langle \log \delta(\mathbf{w}_2; \widehat{\mathbf{w}}_2) \rangle_{\delta(\mathbf{w}_2; \widehat{\mathbf{w}}_2)} \\ &= \left\langle \log \frac{r_1(\mathbf{w}_1)}{p(\mathbf{w}_1|\widehat{\mathbf{w}}_2, \mathcal{D})} \right\rangle_{r_1(\mathbf{w}_1)} - \log p(\mathcal{D}|\widehat{\mathbf{w}}_2)p(\widehat{\mathbf{w}}_2) + \langle \log \delta(\mathbf{w}_2; \widehat{\mathbf{w}}_2) \rangle_{\delta(\mathbf{w}_2; \widehat{\mathbf{w}}_2)}, \end{aligned} \quad (2.39)$$

where

$$p(\mathcal{D}|\widehat{\mathbf{w}}_2) = \langle p(\mathcal{D}|\mathbf{w}_1, \widehat{\mathbf{w}}_2) \rangle_{p(\mathbf{w}_1)} = \int p(\mathcal{D}|\mathbf{w}_1, \widehat{\mathbf{w}}_2)p(\mathbf{w}_1)d\mathbf{w}_1. \quad (2.40)$$

The free energy (2.39) depends on r_1 only through the first term, which is the KL divergence, $\text{KL}(r_1(\mathbf{w}_1)||p(\mathbf{w}_1|\widehat{\mathbf{w}}_2, \mathcal{D}))$, from the trial distribution to the Bayes posterior (conditioned on $\widehat{\mathbf{w}}_2$). Therefore, the minimizer for r_1 is trivially the conditional Bayes posterior

$$r_1(\mathbf{w}_1) = p(\mathbf{w}_1|\widehat{\mathbf{w}}_2, \mathcal{D}), \quad (2.41)$$

with which the first term in Eq. (2.39) vanishes. The third term in Eq. (2.39) is the entropy of the delta function, which diverges to infinity but is independent of $\widehat{\mathbf{w}}_2$. By regarding the delta function as a distribution with its width narrow enough to express a point estimate, while its entropy is finite (although it is very large), we can ignore the third term. Thus, the free energy minimization problem (2.38) can be written as

$$\min_{\widehat{\mathbf{w}}_2} -\log p(\mathcal{D}|\widehat{\mathbf{w}}_2)p(\widehat{\mathbf{w}}_2), \quad (2.42)$$

which amounts to MAP learning for \mathbf{w}_2 after \mathbf{w}_1 is marginalized out.

This method is computationally beneficial when the likelihood $p(\mathcal{D}|\mathbf{w}) = p(\mathcal{D}|\mathbf{w}_1, \mathbf{w}_2)$ is conditionally conjugate to the prior $p(\mathbf{w}_1)$ with respect to \mathbf{w}_1 , given \mathbf{w}_2 . Thanks to the conditional conjugacy, the posterior (2.41) of \mathbf{w}_1 is in a known form, and its normalization factor (2.40), which is required when evaluating the objective in Eq. (2.42), can be obtained analytically.

PB learning was applied in many previous works. For example, in the *expectation-maximization (EM) algorithm* (Dempster et al., 1977), latent variables are integrated out and parameters are point-estimated. In the first probabilistic interpretation of principal component analysis (PCA) (Tipping and Bishop, 1999), one factor of the matrix factorization was called a latent variable and integrated out, while the other factor was called a parameter and point-estimated.

The same idea has been adopted for Gibbs sampling and VB learning, where some of the unknown parameters are integrated out based on the conditional conjugacy, and the other parameters are estimated by the corresponding learning method. Those methods are called *collapsed Gibbs sampling* (Griffiths and Steyvers, 2004) and *collapsed VB learning* (Kurihara et al., 2007; Teh et al., 2007; Sato et al., 2012), respectively. Following this terminology, PB learning may be also called *collapsed MAP learning*. The collapsed version is in general more accurate and more computationally efficient than the uncollapsed counterpart, since it imposes a weaker constraint and applies a nonexact numerical estimation to a smaller number of unknowns.

2.2.3 Expectation Propagation

As explained in Section 2.1.1, VB learning amounts to minimizing the KL divergence $\text{KL}(r(\mathbf{w})\|p(\mathbf{w}|\mathcal{D}))$ from the approximate posterior to the Bayes posterior. *Expectation propagation (EP)* is an alternative deterministic approximation scheme, which minimizes the KL divergence *from the Bayes posterior to the approximate posterior* (Minka, 2001b), i.e.,

$$\min_r \text{KL}(p(\mathbf{w}|\mathcal{D})\|r(\mathbf{w})) \quad \text{s.t.} \quad r \in \mathcal{G}. \quad (2.43)$$

Clearly from its definition, the KL divergence,

$$\text{KL}(q(\mathbf{x})\|p(\mathbf{x})) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x},$$

diverges to $+\infty$ if the support of $q(\mathbf{x})$ is not covered by the support of $p(\mathbf{x})$, while it remains finite if the support of $p(\mathbf{x})$ is not covered by the support of $q(\mathbf{x})$. Due to this asymmetric property of the KL divergence, VB learning and EP can provide drastically different approximate posteriors—VB learning,

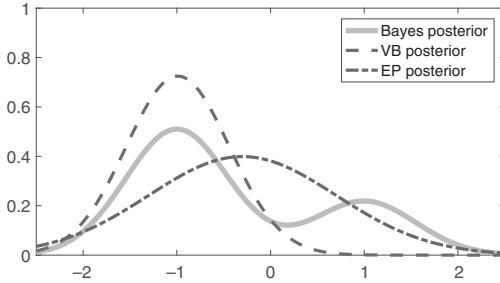


Figure 2.1 Bayes posterior, VB posterior, and EP posterior.

minimizing $\text{KL}(r(\mathbf{w})\|p(\mathbf{w}|\mathcal{D}))$, tends to provide a posterior that approximates a single mode of the Bayes posterior, while EP, minimizing $\text{KL}(p(\mathbf{w}|\mathcal{D})\|r(\mathbf{w}))$, tends to provide a posterior with a broad support covering all modes of the Bayes posterior (see the illustration in Figure 2.1).

Moment Matching Algorithm

The EP problem (2.43) is typically solved by *moment matching*. It starts with expressing the posterior distribution by the product of factors,

$$p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z} \prod_n t_n(\mathbf{w}),$$

where $Z = p(\mathcal{D})$ is the marginal likelihood. For example, in the parametric density estimation (Example 1.1) with i.i.d. samples $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$, the factor can be set to $t_n(\mathbf{w}) = p(\mathbf{x}^{(n)}|\mathbf{w})$ and $t_0(\mathbf{w}) = p(\mathbf{w})$. In EP, the approximating posterior is also assumed to have the same form,

$$r(\mathbf{w}) = \frac{1}{\tilde{Z}} \prod_n \tilde{t}_n(\mathbf{w}), \quad (2.44)$$

where \tilde{Z} is the normalization constant and becomes an approximation of the marginal likelihood Z . Note that the factorization is not over the elements of \mathbf{w} .

EP tries to minimize the KL divergence,

$$\text{KL}(p\|r) = \text{KL}\left(\frac{1}{Z} \prod_n t_n(\mathbf{w}) \parallel \frac{1}{\tilde{Z}} \prod_n \tilde{t}_n(\mathbf{w})\right),$$

which is approximately carried out by refining each factor while the other factors are fixed, and cycling through all the factors. To refine the factor $\tilde{t}_n(\mathbf{w})$, we define the unnormalized distribution,

$$r^{-n}(\mathbf{w}) = \frac{r(\mathbf{w})}{\tilde{t}_n(\mathbf{w})},$$

and the following distribution is used as an estimator of the true posterior:

$$\widehat{p}_n(\mathbf{w}) = \frac{t_n(\mathbf{w})r^{-n}(\mathbf{w})}{\widetilde{Z}_n},$$

where $\widetilde{Z}_n = \int t_n(\mathbf{w})r^{-n}(\mathbf{w})d\mathbf{w}$ is the normalization constant. That is, the new approximating posterior $r^{\text{new}}(\mathbf{w})$ is computed so that it minimizes $\text{KL}(\widehat{p}_n||r^{\text{new}})$. Usually, the approximating posterior is assumed to be a member of the exponential family. In that case, the minimization of $\text{KL}(\widehat{p}_n||r^{\text{new}})$ is reduced to the moment matching between \widehat{p}_n and r^{new} . Namely, the parameter of r^{new} is determined so that its moments are matched with those of \widehat{p}_n .

The new approximating posterior r^{new} yields the refinement of the factor $\widetilde{t}_n(\mathbf{w})$,

$$\widetilde{t}_n(\mathbf{w}) = \widetilde{Z}_n \frac{r^{\text{new}}(\mathbf{w})}{r^{-n}(\mathbf{w})},$$

where the multiplication of \widetilde{Z}_n is derived from the zeroth-order moment matching between \widehat{p}_n and r^{new} , $\int t_n(\mathbf{w})r^{-n}(\mathbf{w})d\mathbf{w} = \int \widetilde{t}_n(\mathbf{w})r^{-n}(\mathbf{w})d\mathbf{w}$.

After several passes through all the factors, if the factors converge, then the posterior is approximated by Eq. (2.44), and the marginal likelihood is approximated by $\widetilde{Z} = \int \prod_n \widetilde{t}_n(\mathbf{w})d\mathbf{w}$ or alternatively by updating it as $\widetilde{Z} \leftarrow \widetilde{Z}\widetilde{Z}_n$ whenever the factor $\widetilde{t}_n(\mathbf{w})$ is refined. Although the convergence of EP is not guaranteed, it is known that if EP converges, the resulting approximating posterior is a stationary point of a certain energy function (Minka, 2001b).

Local Variational Approximation for EP

In Section 2.1.1, we saw that VB learning *minimizes an upper-bound* (the free energy (2.4)) of the Bayes free energy $F^{\text{Bayes}} \equiv -\log p(\mathcal{D})$ (or equivalently maximizing the ELBO). We can say that EP does the opposite. Namely, the EP problem (2.43) *maximizes a lowerbound* of the Bayes free energy:

$$\max_r E(r) \quad \text{s.t.} \quad r \in \mathcal{G}, \quad (2.45)$$

$$\text{where } E(r) = - \int \frac{p(\mathbf{w}, \mathcal{D})}{p(\mathcal{D})} \log \frac{p(\mathbf{w}, \mathcal{D})}{r(\mathbf{w})} d\mathbf{w} \quad (2.46)$$

$$\begin{aligned} &= - \int p(\mathbf{w}|\mathcal{D}) \log \frac{p(\mathbf{w}|\mathcal{D})}{r(\mathbf{w})} d\mathbf{w} - \log p(\mathcal{D}) \\ &= -\text{KL}(p(\mathbf{w}|\mathcal{D})||r(\mathbf{w})) - \log p(\mathcal{D}). \end{aligned} \quad (2.47)$$

The maximization form (2.45) of the EP problem can be solved by local variational approximation, which is akin to the local variational approximation for VB learning (Section 2.1.7). Let us restrict the search space \mathcal{G} for the approximate posterior $r(\mathbf{w}; \widehat{\nu})$ to the function space of a simple distribution

family, e.g., Gaussian, parameterized with variational parameters $\widehat{\boldsymbol{\nu}}$. Then, the EP problem (2.45) is reduced to the following unconstrained maximization problem,

$$\max_{\widehat{\boldsymbol{\nu}}} E(\widehat{\boldsymbol{\nu}}), \quad (2.48)$$

of the objective function written as

$$E(\widehat{\boldsymbol{\nu}}) = - \int \frac{p(\mathbf{w}, \mathcal{D})}{p(\mathcal{D})} \log \frac{p(\mathbf{w}, \mathcal{D})}{p(\mathcal{D})r(\mathbf{w}; \widehat{\boldsymbol{\nu}})} d\mathbf{w} - \log p(\mathcal{D}). \quad (2.49)$$

Consider lower-bounding the objective (2.49) as

$$E(\widehat{\boldsymbol{\nu}}) \geq \underline{E}(\widehat{\boldsymbol{\nu}}, \boldsymbol{\eta}) \equiv - \int \frac{\bar{p}(\mathbf{w}; \boldsymbol{\eta})}{p(\mathcal{D})} \max \left\{ 0, \log \frac{\bar{p}(\mathbf{w}; \boldsymbol{\eta})}{p(\mathcal{D})r(\mathbf{w}; \widehat{\boldsymbol{\nu}})} \right\} d\mathbf{w} - \log p(\mathcal{D}) \quad (2.50)$$

by using a parametric upper-bound $\bar{p}(\mathbf{w}; \boldsymbol{\eta})$ of the joint distribution such that

$$\bar{p}(\mathbf{w}; \boldsymbol{\eta}) \geq p(\mathbf{w}, \mathcal{D}) \quad (2.51)$$

for any $\mathbf{w} \in \mathcal{W}$ and $\boldsymbol{\eta} \in \mathcal{H}$, where $\boldsymbol{\eta}$ is another set of variational parameters with its domain \mathcal{H} .³ Let us choose an upper-bound $\bar{p}(\mathbf{w}; \boldsymbol{\eta})$ such that its function form with respect to \mathbf{w} is the same as the approximate posterior $r(\mathbf{w}; \widehat{\boldsymbol{\nu}})$. More specifically, we assume that, for any given $\boldsymbol{\eta}$, there exists $\widehat{\boldsymbol{\nu}}$ such that

$$\bar{p}(\mathbf{w}; \boldsymbol{\eta}) \propto r(\mathbf{w}; \widehat{\boldsymbol{\nu}}) \quad (2.52)$$

as a function of \mathbf{w} .

Since the direct maximization of $E(\widehat{\boldsymbol{\nu}})$ is intractable, we instead maximize its lower-bound $\underline{E}(\widehat{\boldsymbol{\nu}}, \boldsymbol{\eta})$ jointly over $\widehat{\boldsymbol{\nu}}$ and $\boldsymbol{\eta}$. Namely, we solve the problem,

$$\max_{\widehat{\boldsymbol{\nu}}, \boldsymbol{\eta}} \underline{E}(\widehat{\boldsymbol{\nu}}, \boldsymbol{\eta}), \quad (2.53)$$

to find the approximate posterior $r(\mathbf{w}; \widehat{\boldsymbol{\nu}})$ such that $\underline{E}(\widehat{\boldsymbol{\nu}}, \boldsymbol{\eta}) (\leq E(\widehat{\boldsymbol{\nu}}))$ is closest to the Bayes free energy F^{Bayes} (when $\boldsymbol{\eta}$ is also optimized).

Let

$$q(\mathbf{w}; \boldsymbol{\eta}) = \frac{\bar{p}(\mathbf{w}; \boldsymbol{\eta})}{\bar{Z}(\boldsymbol{\eta})} \quad (2.54)$$

be the distribution created by normalizing the upper-bound with its normalization factor

$$\bar{Z}(\boldsymbol{\eta}) = \int \bar{p}(\mathbf{w}; \boldsymbol{\eta}) d\mathbf{w}. \quad (2.55)$$

³ The two sets, $\widehat{\boldsymbol{\nu}}$ and $\boldsymbol{\eta}$, of variational parameters play the same roles as $\widehat{\lambda}$ and ξ , respectively, in the local variational approximation for VB learning.

Note that the normalization factor (2.55) is trivially an upper-bound of the marginal likelihood, i.e.,

$$\bar{Z}(\boldsymbol{\eta}) \geq \int p(\mathbf{w}, \mathcal{D}) d\mathbf{w} = p(\mathcal{D}),$$

and is tractable because of the assumption (2.52) that \bar{p} is in the same simple function form as r .

With Eq. (2.54), the lower-bound (2.50) is expressed as

$$\begin{aligned} \underline{E}(\widehat{\boldsymbol{\nu}}, \boldsymbol{\eta}) &= - \int \frac{\bar{Z}(\boldsymbol{\eta}) q(\mathbf{w}; \boldsymbol{\eta})}{p(\mathcal{D})} \max \left\{ 0, \log \frac{\bar{Z}(\boldsymbol{\eta}) q(\mathbf{w}; \boldsymbol{\eta})}{p(\mathcal{D}) r(\mathbf{w}; \widehat{\boldsymbol{\nu}})} \right\} d\mathbf{w} - \log p(\mathcal{D}) \\ &= - \frac{\bar{Z}(\boldsymbol{\eta})}{p(\mathcal{D})} \int q(\mathbf{w}; \boldsymbol{\eta}) \max \left\{ 0, \log \frac{q(\mathbf{w}; \boldsymbol{\eta})}{r(\mathbf{w}; \widehat{\boldsymbol{\nu}})} + \log \frac{\bar{Z}(\boldsymbol{\eta})}{p(\mathcal{D})} \right\} d\mathbf{w} - \log p(\mathcal{D}). \end{aligned} \quad (2.56)$$

Eq. (2.56) is upper-bounded by

$$\begin{aligned} &- \frac{\bar{Z}(\boldsymbol{\eta})}{p(\mathcal{D})} \int q(\mathbf{w}; \boldsymbol{\eta}) \left(\log \frac{q(\mathbf{w}; \boldsymbol{\eta})}{r(\mathbf{w}; \widehat{\boldsymbol{\nu}})} + \log \frac{\bar{Z}(\boldsymbol{\eta})}{p(\mathcal{D})} \right) d\mathbf{w} - \log p(\mathcal{D}) \\ &= - \frac{\bar{Z}(\boldsymbol{\eta})}{p(\mathcal{D})} \text{KL}(q(\mathbf{w}; \boldsymbol{\eta}) \| r(\mathbf{w}; \widehat{\boldsymbol{\nu}})) - \frac{\bar{Z}(\boldsymbol{\eta})}{p(\mathcal{D})} \log \frac{\bar{Z}(\boldsymbol{\eta})}{p(\mathcal{D})} - \log p(\mathcal{D}), \end{aligned} \quad (2.57)$$

which, for any $\boldsymbol{\eta} \in \mathbf{H}$, is maximized when $\widehat{\boldsymbol{\nu}}$ is such that

$$r(\mathbf{w}; \widehat{\boldsymbol{\nu}}) = q(\mathbf{w}; \boldsymbol{\eta}) \quad (2.58)$$

(the assumption (2.52) guarantees the attainability). With this optimal $\widehat{\boldsymbol{\nu}}$, Eq. (2.57) coincides with Eq. (2.56). Thus, after optimization with respect to $\widehat{\boldsymbol{\nu}}$, the lower-bound (2.56) is given as

$$\max_{\widehat{\boldsymbol{\nu}}} \underline{E}(\widehat{\boldsymbol{\nu}}, \boldsymbol{\eta}) = - \frac{\bar{Z}(\boldsymbol{\eta})}{p(\mathcal{D})} \log \frac{\bar{Z}(\boldsymbol{\eta})}{p(\mathcal{D})} - \log p(\mathcal{D}). \quad (2.59)$$

Since $x \log x$ for $x \geq 1$ is monotonically increasing, maximizing the lower-bound (2.59) is achieved by solving

$$\min_{\boldsymbol{\eta}} \bar{Z}(\boldsymbol{\eta}). \quad (2.60)$$

Once the minimizer $\widehat{\boldsymbol{\eta}}$ is obtained, Eq. (2.58) gives the optimal approximate posterior.

The problem (2.60) amounts to minimizing an upper-bound of the marginal likelihood. This is in contrast to the local variational approximation for VB learning, where a lower-bound of the marginal likelihood is maximized in the end (compare Eq. (2.33) and Eq. (2.60)).

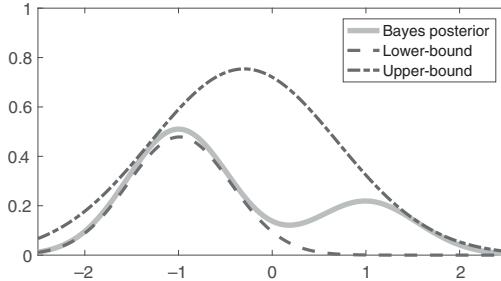


Figure 2.2 Bayes posterior and its tightest lower- and upper-bounds, formed by a Gaussian.

Remembering that the joint distribution is proportional to the Bayes posterior, i.e., $p(\mathbf{w}|\mathcal{D}) = p(\mathbf{w}, \mathcal{D})/p(\mathcal{D})$, we can say that the VB posterior is the normalized version of the tightest (in terms of the total mass) lower-bound of the Bayes posterior, while the EP posterior is the normalized version of the tightest upper-bound of the Bayes posterior. Figure 2.2 illustrates the tightest upper-bound and the tightest lower-bound of the Bayes posterior, which correspond to unnormalized versions of the VB posterior and the EP posterior, respectively (compare Figures 2.1 and 2.2). This view also explains the tendency of VB learning and EP—a lower-bound (the VB posterior) must be zero wherever the Bayes posterior is zero, while an upper-bound (the EP posterior) must be positive wherever the Bayes posterior is positive.

2.2.4 Metropolis–Hastings Sampling

If a sufficient number of samples $\{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(L)}\}$ from the posterior distribution (1.3) are obtained, the expectation $\int f(\mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w}$ required for computing the quantities such as Eqs. (1.6) through (1.9) can be approximated by

$$\frac{1}{L} \sum_{l=1}^L f(\mathbf{w}^{(l)}).$$

The *Metropolis–Hastings sampling* and the Gibbs sampling are most popular methods to sample from the (unnormalized) posterior distribution in the framework of *Markov chain Monte Carlo (MCMC)*.

In the Metropolis–Hastings sampling, we draw samples from a simple distribution $q(\mathbf{w}|\mathbf{w}^{(t)})$ called a proposal distribution, which is conditioned on the current state $\mathbf{w}^{(t)}$ of the parameter (or latent variables) \mathbf{w} . The proposal distribution is chosen to be a simple distribution such as a Gaussian centered at $\mathbf{w}^{(t)}$ if \mathbf{w} is continuous or the uniform distribution in a certain neighborhood

of $\mathbf{w}^{(t)}$ if \mathbf{w} is discrete. At each cycle of the algorithm, we draw a candidate sample \mathbf{w}^* from the proposal distribution $q(\mathbf{w}|\mathbf{w}^{(t)})$, and we accept it with probability

$$\min \left(1, \frac{p(\mathbf{w}^*, \mathcal{D})}{p(\mathbf{w}^{(t)}, \mathcal{D})} \frac{q(\mathbf{w}^{(t)}|\mathbf{w}^*)}{q(\mathbf{w}^*|\mathbf{w}^{(t)})} \right).$$

If \mathbf{w}^* is accepted, then the next state $\mathbf{w}^{(t+1)}$ is moved to \mathbf{w}^* , $\mathbf{w}^{(t+1)} = \mathbf{w}^*$; otherwise, it stays at the current state, $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)}$. We repeat this procedure until a sufficiently long sequence of states is obtained. Note that if the proposal distribution is symmetric, i.e., $q(\mathbf{w}|\mathbf{w}') = q(\mathbf{w}'|\mathbf{w})$ for any \mathbf{w} and \mathbf{w}' , in which case the algorithm is called the Metropolis algorithm, the probability of acceptance depends on the ratio of the posteriors,

$$\frac{p(\mathbf{w}^*, \mathcal{D})}{p(\mathbf{w}^{(t)}, \mathcal{D})} = \frac{p(\mathbf{w}^*, \mathcal{D})/Z}{p(\mathbf{w}^{(t)}, \mathcal{D})/Z} = \frac{p(\mathbf{w}^*|\mathcal{D})}{p(\mathbf{w}^{(t)}|\mathcal{D})},$$

and if \mathbf{w}^* has higher posterior probability (density) than $\mathbf{w}^{(t)}$, it is accepted with probability 1.

To guarantee that the distribution of the sampled sequence converges to the posterior distribution, we discard a first part of the sequence, which is called *burn-in*. Usually, after the burn-in period, we retain only every M th sample and discard the other samples so that the retained samples can be considered as independent if M is sufficiently large.

2.2.5 Gibbs Sampling

Another popular MCMC method is *Gibbs sampling*, which makes use of the conditional conjugacy. More specifically, it is applicable when we can compute and draw samples from the conditional distribution of a variable of $\mathbf{w} \in \mathbb{R}^J$,

$$p(w_j|w_1, \dots, w_{j-1}, w_{j+1}, \dots, w_J, \mathcal{D}) \equiv p(w_j|\mathbf{w}_{\neg j}, \mathcal{D}),$$

conditioned on the rest of the variables of \mathbf{w} .

Assuming that $\mathbf{w}^{(t)}$ is obtained at the t th cycle of the Gibbs sampling algorithm, the next sample of each variable is drawn from the conditional distribution,

$$p(w_j^{(t+1)}|\mathbf{w}_{\neg j}^{(t)}, \mathcal{D}),$$

where

$$\mathbf{w}_{\neg j}^{(t)} = (w_1^{(t+1)}, \dots, w_{j-1}^{(t+1)}, w_{j+1}^{(t)}, \dots, w_J^{(t)})$$

from $j = 1$ to J in turn.

This sampling procedure can be viewed as a special case of the Metropolis–Hastings algorithm. If the proposal distribution $q(\mathbf{w}|\mathbf{w}^{(t)})$ is chosen to be

$$p(\mathbf{w}_j | \mathbf{w}_{\neg j}^{(t)}, \mathcal{D}) \delta(\mathbf{w}_{\neg j} - \mathbf{w}_{\neg j}^{(t)}),$$

then the probability that the candidate \mathbf{w}^* is accepted is 1 since $\mathbf{w}_{\neg j}^* = \mathbf{w}_{\neg j}^{(t)}$ implies that

$$\begin{aligned} \frac{p(\mathbf{w}^*, \mathcal{D})}{p(\mathbf{w}^{(t)}, \mathcal{D})} \frac{q(\mathbf{w}^{(t)} | \mathbf{w}^*)}{q(\mathbf{w}^* | \mathbf{w}^{(t)})} &= \frac{p(\mathbf{w}^* | \mathcal{D})}{p(\mathbf{w}^{(t)} | \mathcal{D})} \frac{p(\mathbf{w}_j^{(t)} | \mathbf{w}_{\neg j}^{(t)}, \mathcal{D})}{p(\mathbf{w}_j^* | \mathbf{w}_{\neg j}^{(t)}, \mathcal{D})} \\ &= \frac{p(\mathbf{w}_{\neg j}^* | \mathcal{D}) p(\mathbf{w}_j^* | \mathbf{w}_{\neg j}^*, \mathcal{D})}{p(\mathbf{w}_{\neg j}^{(t)} | \mathcal{D}) p(\mathbf{w}_j^{(t)} | \mathbf{w}_{\neg j}^{(t)}, \mathcal{D})} \frac{p(\mathbf{w}_j^{(t)} | \mathbf{w}_{\neg j}^{(t)}, \mathcal{D})}{p(\mathbf{w}_j^* | \mathbf{w}_{\neg j}^{(t)}, \mathcal{D})} = 1. \end{aligned}$$

As we have seen in the Metropolis–Hastings and Gibbs sampling algorithms, MCMC methods do not require the knowledge of the normalization constant $Z = \int p(\mathbf{w}, \mathcal{D}) d\mathbf{w}$. Note that, however, even if we have samples from the posterior, we need additional steps to compute Z with the samples. A simple way is to calculate the expectation of the inverse of the likelihood by the sample average,

$$\left\langle \frac{1}{p(\mathcal{D}|\mathbf{w})} \right\rangle_{p(\mathbf{w}|\mathcal{D})} \approx \frac{1}{L} \sum_{l=1}^L \frac{1}{p(\mathcal{D}|\mathbf{w}^{(l)})}.$$

It provides an estimate of the inverse of Z because

$$\left\langle \frac{1}{p(\mathcal{D}|\mathbf{w})} \right\rangle_{p(\mathbf{w}|\mathcal{D})} = \int \frac{1}{p(\mathcal{D}|\mathbf{w})} \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{Z} d\mathbf{w} = \frac{1}{Z} \int p(\mathbf{w}) d\mathbf{w} = \frac{1}{Z}.$$

However, this estimator is known to have high variance. A more sophisticated sampling method to compute Z was developed by Chib (1995), while it requires multiple runs of MCMC sampling.

A new efficient method to compute the marginal likelihood was recently proposed and named a *widely applicable Bayesian information criterion (WBIC)*, which requires only a single run of MCMC sampling from a generalized posterior distribution (Watanabe, 2013). This method computes the expectation of the negative log-likelihood,

$$\langle -\log p(\mathcal{D}|\mathbf{w}) \rangle_{p^{(\beta)}(\mathbf{w}|\mathcal{D})},$$

over the β -generalized posterior distribution defined as

$$p^{(\beta)}(\mathbf{w}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{w})^\beta p(\mathbf{w}),$$

with $\beta = 1/\log N$, where N is the number of i.i.d. samples. The computed (approximated) expectation is proved to have the same leading terms as those of the asymptotic expansion of $-\log Z$ as $N \rightarrow \infty$.

Part II

Algorithm

3

VB Algorithm for Multilinear Models

In this chapter, we derive iterative VB algorithms for multilinear models with Gaussian noise, where we can rely on the conditional conjugacy with respect to each linear factor. The models introduced in this chapter will be further analyzed in Part III, where the global solution or its approximation is analytically derived, and the behavior of the VB solution is investigated in detail.

3.1 Matrix Factorization

Assume that we observe a matrix $\mathbf{V} \in \mathbb{R}^{L \times M}$, which is the sum of a target matrix $\mathbf{U} \in \mathbb{R}^{L \times M}$ and a noise matrix $\mathcal{E} \in \mathbb{R}^{L \times M}$:

$$\mathbf{V} = \mathbf{U} + \mathcal{E}.$$

In the *matrix factorization (MF)* model (Srebro and Jaakkola, 2003; Srebro et al., 2005; Lim and Teh, 2007; Salakhutdinov and Mnih, 2008; Ilin and Raiko, 2010) or the *probabilistic principal component analysis (probabilistic PCA)* (Tipping and Bishop, 1999; Bishop, 1999b), the target matrix is assumed to be low rank, and therefore can be factorized as

$$\mathbf{U} = \mathbf{B}\mathbf{A}^\top,$$

where $\mathbf{A} \in \mathbb{R}^{M \times H}$, $\mathbf{B} \in \mathbb{R}^{L \times H}$ for $H \leq \min(L, M)$ are unknown parameters to be estimated, and \top denotes the transpose of a matrix or vector. Here, the rank of \mathbf{U} is upper-bounded by H . We denote a column vector of a matrix by a bold lowercase letter, and a row vector by a bold lowercase letter with a tilde, namely,

$$\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_H) = (\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_M)^\top \in \mathbb{R}^{M \times H},$$

$$\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_H) = (\tilde{\mathbf{b}}_1, \dots, \tilde{\mathbf{b}}_L)^\top \in \mathbb{R}^{L \times H}.$$

3.1.1 VB Learning for MF

Assume that the observation noise \mathcal{E} is independent Gaussian:

$$p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{V} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2\right), \quad (3.1)$$

where $\|\cdot\|_{\text{Fro}}$ denotes the Frobenius norm.

Conditional Conjugacy

If we treat \mathbf{B} as a constant, the likelihood (3.1) is in the Gaussian form of \mathbf{A} . Similarly, if we treat \mathbf{A} as a constant, the likelihood (3.1) is in the Gaussian form of \mathbf{B} . Therefore, conditional conjugacy with respect to \mathbf{A} given \mathbf{B} , as well as with respect to \mathbf{B} given \mathbf{A} , holds if we adopt Gaussian priors:

$$p(\mathbf{A}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{A}\mathbf{C}_A^{-1}\mathbf{A}^\top)\right), \quad (3.2)$$

$$p(\mathbf{B}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{B}\mathbf{C}_B^{-1}\mathbf{B}^\top)\right), \quad (3.3)$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix.

Typically, the prior covariance matrices \mathbf{C}_A and \mathbf{C}_B are restricted to be diagonal, which induces low-rankness (we discuss this mechanism in Chapter 7):

$$\begin{aligned} \mathbf{C}_A &= \mathbf{Diag}(c_{a_1}^2, \dots, c_{a_H}^2), \\ \mathbf{C}_B &= \mathbf{Diag}(c_{b_1}^2, \dots, c_{b_H}^2), \end{aligned}$$

for $c_{a_h}, c_{b_h} > 0, h = 1, \dots, H$.

Variational Bayesian Algorithm

Thanks to the conditional conjugacy, the following independence constraint makes the approximate posterior Gaussian:

$$r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}). \quad (3.4)$$

The VB learning problem (2.13) is then reduced to

$$\widehat{r} = \underset{r}{\operatorname{argmin}} F(r) \quad \text{s.t.} \quad r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}). \quad (3.5)$$

Under the constraint (3.4), the free energy is written as

$$\begin{aligned} F(r) &= \left\langle \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{p(\mathbf{V}|\mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\ &= \int r_A(\mathbf{A})r_B(\mathbf{B}) \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{p(\mathbf{V}|\mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})} d\mathbf{A}d\mathbf{B}. \end{aligned} \quad (3.6)$$

Following the recipe described in Section 2.1.5, we take the derivatives of the free energy (3.6) with respect to $r_A(\mathbf{A})$ and $r_B(\mathbf{B})$, respectively. Thus, we obtain the following stationary conditions:

$$r_A(\mathbf{A}) \propto p(\mathbf{A}) \exp \langle \log p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \rangle_{r_B(\mathbf{B})}, \quad (3.7)$$

$$r_B(\mathbf{B}) \propto p(\mathbf{B}) \exp \langle \log p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \rangle_{r_A(\mathbf{A})}. \quad (3.8)$$

By substituting the likelihood (3.1) and the prior (3.2) into Eq. (3.7), we obtain

$$\begin{aligned} r_A(\mathbf{A}) &\propto \exp \left(-\frac{1}{2} \text{tr}(\mathbf{A} \mathbf{C}_A^{-1} \mathbf{A}^\top) - \frac{1}{2\sigma^2} \left\langle \|\mathbf{V} - \mathbf{B} \mathbf{A}^\top\|_{\text{Fro}}^2 \right\rangle_{r_B(\mathbf{B})} \right) \\ &\propto \exp \left(-\frac{1}{2} \text{tr}(\mathbf{A} \mathbf{C}_A^{-1} \mathbf{A}^\top + \sigma^{-2} \langle -2\mathbf{V}^\top \mathbf{B} \mathbf{A}^\top + \mathbf{A} \mathbf{B}^\top \mathbf{B} \mathbf{A}^\top \rangle_{r_B(\mathbf{B})}) \right) \\ &\propto \exp \left(-\frac{\text{tr}((\mathbf{A} - \widehat{\mathbf{A}}) \widehat{\Sigma}_A^{-1} (\mathbf{A} - \widehat{\mathbf{A}})^\top)}{2} \right), \end{aligned} \quad (3.9)$$

where

$$\widehat{\mathbf{A}} = \sigma^{-2} \mathbf{V}^\top \langle \mathbf{B} \rangle_{r_B(\mathbf{B})} \widehat{\Sigma}_A, \quad (3.10)$$

$$\widehat{\Sigma}_A = \sigma^2 \left(\langle \mathbf{B}^\top \mathbf{B} \rangle_{r_B(\mathbf{B})} + \sigma^2 \mathbf{C}_A^{-1} \right)^{-1}. \quad (3.11)$$

Similarly, by substituting the likelihood (3.1) and the prior (3.3) into Eq. (3.8), we obtain

$$\begin{aligned} r_B(\mathbf{B}) &\propto \exp \left(-\frac{1}{2} \text{tr}(\mathbf{B} \mathbf{C}_B^{-1} \mathbf{B}^\top) - \frac{1}{2\sigma^2} \left\langle \|\mathbf{V} - \mathbf{B} \mathbf{A}^\top\|_{\text{Fro}}^2 \right\rangle_{r_A(\mathbf{A})} \right) \\ &\propto \exp \left(-\frac{1}{2} \text{tr}(\mathbf{B} \mathbf{C}_B^{-1} \mathbf{B}^\top + \sigma^{-2} \langle -2\mathbf{V} \mathbf{A} \mathbf{B}^\top + \mathbf{B} \mathbf{A}^\top \mathbf{A} \mathbf{B}^\top \rangle_{r_A(\mathbf{A})}) \right) \\ &\propto \exp \left(-\frac{\text{tr}((\mathbf{B} - \widehat{\mathbf{B}}) \widehat{\Sigma}_B^{-1} (\mathbf{B} - \widehat{\mathbf{B}})^\top)}{2} \right), \end{aligned} \quad (3.12)$$

where

$$\widehat{\mathbf{B}} = \sigma^{-2} \mathbf{V} \langle \mathbf{A} \rangle_{r_A(\mathbf{A})} \widehat{\Sigma}_B, \quad (3.13)$$

$$\widehat{\Sigma}_B = \sigma^2 \left(\langle \mathbf{A}^\top \mathbf{A} \rangle_{r_A(\mathbf{A})} + \sigma^2 \mathbf{C}_B^{-1} \right)^{-1}. \quad (3.14)$$

Eqs. (3.9) and (3.12) imply that the posteriors are Gaussian. More specifically, they can be written as

$$r_A(\mathbf{A}) = \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \mathbf{I}_M \otimes \widehat{\Sigma}_A), \quad (3.15)$$

$$r_B(\mathbf{B}) = \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \mathbf{I}_L \otimes \widehat{\Sigma}_B), \quad (3.16)$$

where \otimes denotes the *Kronecker product*, and

$$\text{MGauss}_{D_1, D_2}(X; M, \check{\Sigma}) \equiv \text{Gauss}_{D_1 \cdot D_2}(\text{vec}(X^\top); \text{vec}(M^\top), \check{\Sigma}) \quad (3.17)$$

denotes the *matrix variate Gaussian distribution* (Gupta and Nagar, 1999). Here, $\text{vec} : \mathbb{R}^{D_2 \times D_1} \mapsto \mathbb{R}^{D_2 D_1}$ is the *vectorization operator*, which concatenates all column vectors of a matrix into a long column vector. Note that, if the covariance has a specific structure expressed as $\check{\Sigma} = \Sigma \otimes \Psi \in \mathbb{R}^{D_2 D_1 \times D_2 D_1}$, such as Eqs. (3.15) and (3.16), the matrix variate Gaussian distribution can be written as

$$\begin{aligned} \text{MGauss}_{D_1, D_2}(X; M, \Sigma \otimes \Psi) &\equiv \frac{1}{(2\pi)^{D_1 D_2 / 2} \det(\Sigma)^{D_2 / 2} \det(\Psi)^{D_1 / 2}} \\ &\cdot \exp\left(-\frac{1}{2} \text{tr}\left(\Sigma^{-1}(X - M)\Psi^{-1}(X - M)^\top\right)\right). \end{aligned} \quad (3.18)$$

The fact that the posterior is Gaussian is a consequence of the forced independence between A and B and conditional conjugacy. The parameters, $\{\widehat{A}, \widehat{B}, \widehat{\Sigma}_A, \widehat{\Sigma}_B\}$, defining the VB posterior (3.15) and (3.16), are the *variational parameters*.

Since $r_A(A)$ and $r_B(B)$ are Gaussian, the first and the (noncentered) second moments can be expressed with variational parameters as follows:

$$\begin{aligned} \langle A \rangle_{r_A(A)} &= \widehat{A}, \\ \langle A^\top A \rangle_{r_A(A)} &= \widehat{A}^\top \widehat{A} + M \widehat{\Sigma}_A, \\ \langle B \rangle_{r_B(B)} &= \widehat{B}, \\ \langle B^\top B \rangle_{r_B(B)} &= \widehat{B}^\top \widehat{B} + L \widehat{\Sigma}_B. \end{aligned}$$

By substituting the preceding into Eqs. (3.10), (3.11), (3.13), and (3.14), we have the following relations among the variational parameters:

$$\widehat{A} = \sigma^{-2} V^\top \widehat{B} \widehat{\Sigma}_A, \quad (3.19)$$

$$\widehat{\Sigma}_A = \sigma^2 \left(\widehat{B}^\top \widehat{B} + L \widehat{\Sigma}_B + \sigma^2 C_A^{-1} \right)^{-1}, \quad (3.20)$$

$$\widehat{B} = \sigma^{-2} V \widehat{A} \widehat{\Sigma}_B, \quad (3.21)$$

$$\widehat{\Sigma}_B = \sigma^2 \left(\widehat{A}^\top \widehat{A} + M \widehat{\Sigma}_A + \sigma^2 C_B^{-1} \right)^{-1}. \quad (3.22)$$

As we see shortly, Eqs. (3.19) through (3.22) are stationary conditions for variational parameters, which can be used as update rules for *coordinate descent* local search (Bishop, 1999b).

Free Energy as a Function of Variational Parameters

By substituting Eqs. (3.15) and (3.16) into Eq. (3.6), we can explicitly write down the free energy as (not a functional but) a function of the unknown variational parameters $\{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B\}$:

$$\begin{aligned}
2F &= 2 \left\langle \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{p(\mathbf{V}|\mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\
&= 2 \left\langle \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} - 2 \langle \log p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\
&= \left\langle M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + L \log \frac{\det(\mathbf{C}_B)}{\det(\widehat{\Sigma}_B)} + \text{tr}(\mathbf{C}_A^{-1}\mathbf{A}^\top \mathbf{A} + \mathbf{C}_B^{-1}\mathbf{B}^\top \mathbf{B}) \right. \\
&\quad \left. - \text{tr}(\widehat{\Sigma}_A^{-1}(\mathbf{A} - \widehat{\mathbf{A}})^\top (\mathbf{A} - \widehat{\mathbf{A}}) + \widehat{\Sigma}_B^{-1}(\mathbf{B} - \widehat{\mathbf{B}})^\top (\mathbf{B} - \widehat{\mathbf{B}})) \right. \\
&\quad \left. + LM \log(2\pi\sigma^2) + \frac{\|\mathbf{V} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2}{\sigma^2} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\
&= M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + L \log \frac{\det(\mathbf{C}_B)}{\det(\widehat{\Sigma}_B)} - \text{tr}(M\widehat{\Sigma}_A^{-1}\widehat{\Sigma}_A + L\widehat{\Sigma}_B^{-1}\widehat{\Sigma}_B) \\
&\quad + \text{tr}(\mathbf{C}_A^{-1}(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M\widehat{\Sigma}_A) + \mathbf{C}_B^{-1}(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L\widehat{\Sigma}_B)) \\
&\quad + LM \log(2\pi\sigma^2) + \left\langle \frac{\|(\mathbf{V} - \widehat{\mathbf{B}}\mathbf{A}^\top) + (\widehat{\mathbf{B}}\mathbf{A}^\top - \mathbf{B}\mathbf{A}^\top)\|_{\text{Fro}}^2}{\sigma^2} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\
&= M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + L \log \frac{\det(\mathbf{C}_B)}{\det(\widehat{\Sigma}_B)} - (L+M)H \\
&\quad + \text{tr}(\mathbf{C}_A^{-1}(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M\widehat{\Sigma}_A) + \mathbf{C}_B^{-1}(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L\widehat{\Sigma}_B)) \\
&\quad + LM \log(2\pi\sigma^2) + \frac{\|\mathbf{V} - \widehat{\mathbf{B}}\mathbf{A}^\top\|_{\text{Fro}}^2}{\sigma^2} + \left\langle \frac{\|\widehat{\mathbf{B}}\mathbf{A}^\top - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2}{\sigma^2} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\
&= LM \log(2\pi\sigma^2) + \frac{\|\mathbf{V} - \widehat{\mathbf{B}}\mathbf{A}^\top\|_{\text{Fro}}^2}{\sigma^2} + M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + L \log \frac{\det(\mathbf{C}_B)}{\det(\widehat{\Sigma}_B)} \\
&\quad - (L+M)H + \text{tr}(\mathbf{C}_A^{-1}(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M\widehat{\Sigma}_A) + \mathbf{C}_B^{-1}(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L\widehat{\Sigma}_B)) \\
&\quad + \sigma^{-2} \left(-\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + (\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M\widehat{\Sigma}_A)(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L\widehat{\Sigma}_B) \right) \}. \tag{3.23}
\end{aligned}$$

Now, the VB learning problem is reduced from the function *optimization* (3.5) to the following *variable optimization*:

$$\begin{aligned} \text{Given } & \mathbf{C}_A, \mathbf{C}_B \in \mathbb{D}_{++}^H, \quad \sigma^2 \in \mathbb{R}_{++}, \\ \min_{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B} & F, \\ \text{s.t. } & \widehat{\mathbf{A}} \in \mathbb{R}^{M \times H}, \widehat{\mathbf{B}} \in \mathbb{R}^{L \times H}, \quad \widehat{\Sigma}_A, \widehat{\Sigma}_B \in \mathbb{S}_{++}^H, \end{aligned} \quad (3.24)$$

where \mathbb{R}_{++} is the set of positive real numbers, \mathbb{S}_{++}^D is the set of $D \times D$ (symmetric) positive definite matrices, and \mathbb{D}_{++}^D is the set of $D \times D$ positive definite diagonal matrices.

We note the following:

- Once the solution $\{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B\}$ of the problem (3.24) is obtained, Eqs. (3.15) and (3.16) specify the VB posterior $\widehat{r}(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B})$.
- We treated the prior covariances \mathbf{C}_A and \mathbf{C}_B and the noise variance σ^2 as hyperparameters, and therefore assumed to be given when the VB problem was solved. However, they can be estimated through the empirical Bayesian procedure, which is explained shortly. They can also be treated as random variables, and their VB posterior can be computed by adopting conjugate Gamma priors and minimizing the free energy under an appropriate independence constraint.
- Eqs. (3.19) through (3.22) coincide with the stationary conditions of the free energy (3.23), which are derived from the derivatives with respect to $\widehat{\mathbf{A}}, \widehat{\Sigma}_A, \widehat{\mathbf{B}}$, and $\widehat{\Sigma}_B$, respectively. Therefore, iterating Eqs. (3.19) through (3.22) gives a local solution to the problem (3.24).

Empirical Variational Bayesian Algorithm

The empirical variational Bayesian (EVB) procedure can be performed by minimizing the free energy also with respect to the hyperparameters:

$$\begin{aligned} \min_{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B, \mathbf{C}_A, \mathbf{C}_B, \sigma^2} & F, \\ \text{s.t. } & \widehat{\mathbf{A}} \in \mathbb{R}^{M \times H}, \widehat{\mathbf{B}} \in \mathbb{R}^{L \times H}, \quad \widehat{\Sigma}_A, \widehat{\Sigma}_B \in \mathbb{S}_{++}^H, \\ & \mathbf{C}_A, \mathbf{C}_B \in \mathbb{D}_{++}^H, \quad \sigma^2 \in \mathbb{R}_{++}. \end{aligned} \quad (3.25)$$

By differentiating the free energy (3.23) with respect to each entry of \mathbf{C}_A and \mathbf{C}_B , we have, for $h = 1, \dots, H$,

$$c_{a_h}^2 = \|\widehat{\mathbf{a}}_h\|^2 / M + (\widehat{\Sigma}_A)_{h,h}, \quad (3.26)$$

$$c_{b_h}^2 = \|\widehat{\mathbf{b}}_h\|^2 / L + (\widehat{\Sigma}_B)_{h,h}. \quad (3.27)$$

Algorithm 1 EVB learning for matrix factorization.

-
- 1: Initialize the variational parameters $(\widehat{\mathbf{A}}, \widehat{\boldsymbol{\Sigma}}_A, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\Sigma}}_B)$, and the hyperparameters $(\mathbf{C}_A, \mathbf{C}_B, \sigma^2)$, for example, $\widehat{\mathbf{A}}_{m,h}, \widehat{\mathbf{B}}_{l,h} \sim \text{Gauss}_1(0, \tau)$, $\widehat{\boldsymbol{\Sigma}}_A = \widehat{\boldsymbol{\Sigma}}_B = \mathbf{C}_A = \mathbf{C}_B = \tau \mathbf{I}_H$, and $\sigma^2 = \tau^2$ for $\tau^2 = \|\mathbf{V}\|_{\text{Fro}}^2 / (LM)$.
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (3.20), (3.19), (3.22), and (3.21) to update $\widehat{\boldsymbol{\Sigma}}_A$, $\widehat{\mathbf{A}}$, $\widehat{\boldsymbol{\Sigma}}_B$, and $\widehat{\mathbf{B}}$, respectively.
 - 3: Apply Eqs. (3.26) and (3.27) for all $h = 1, \dots, H$, and Eq. (3.28) to update \mathbf{C}_A , \mathbf{C}_B , and σ^2 , respectively.
 - 4: Prune the h th component if $c_{a_h}^2 c_{b_h}^2 < \varepsilon$, where $\varepsilon > 0$ is a small threshold, e.g., set to $\varepsilon = 10^{-4}$.
 - 5: Evaluate the free energy (3.23).
 - 6: Iterate Steps 2 through 5 until convergence (until the energy decrease becomes smaller than a threshold).
-

Similarly, by differentiating the free energy (3.23) with respect to σ^2 , we have

$$\sigma^2 = \frac{\|\mathbf{V}\|_{\text{Fro}}^2 - \text{tr}(\mathbf{V}^\top \widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top) + \text{tr}((\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\boldsymbol{\Sigma}}_A)(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\boldsymbol{\Sigma}}_B))}{LM}. \quad (3.28)$$

Eqs. (3.26)–(3.28) are used as update rules for the prior covariances \mathbf{C}_A , \mathbf{C}_B , and the noise variance σ^2 , respectively.

Starting from some initial value, iterating Eqs. (3.19) through (3.22) and Eqs. (3.26) through (3.28) gives a local solution for EVB learning. Algorithm 1 summarizes this iterative procedure. If we appropriately set the hyperparameters $(\mathbf{C}_A, \mathbf{C}_B, \sigma^2)$ in Step 1 and skip Steps 3 and 4, Algorithm 1 is reduced to (nonempirical) VB learning.

We note the following for implementation:

- Due to the automatic relevance determination (ARD) effect in EVB learning (see Chapter 7), $c_{a_h}^2 c_{b_h}^2$ converges to zero for some h . For this reason, “pruning” in Step 4 is necessary for numerical stability ($\log \det(\mathbf{C})$ diverges if \mathbf{C} is singular). If the h th component is pruned, the corresponding h th column of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ and the h th column and row of $\widehat{\boldsymbol{\Sigma}}_A$, $\widehat{\boldsymbol{\Sigma}}_B$, \mathbf{C}_A , \mathbf{C}_B should be removed, and the rank H should be reduced accordingly.
- In principle, the update rules never increase the free energy. However, pruning can slightly increase it.
- When computing the free energy by Eq. (3.23), $\log \det(\cdot)$ should be computed as twice the sum of the log of the diagonals of the Cholesky decomposition, i.e.,

$$\log \det(\mathbf{C}) = 2 \sum_{h=1}^H (\log(\text{Chol}(\mathbf{C}))_{h,h}).$$

Otherwise, $\det(\cdot)$ can be huge for practical size of H , causing numerical instability.

Simple Variational Bayesian Learning (with Columnwise Independence)

The updates (3.19) through (3.22) require inversion of an $H \times H$ matrix. One can derive a faster VB learning algorithm by using a stronger constraint for the VB learning. More specifically, instead of the matrixwise independence (3.4), we assume the independence between the column vectors of $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_H)$ and $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_H)$ (Ilin and Raiko, 2010; Nakajima and Sugiyama, 2011; Kim and Choi, 2014):

$$r(\mathbf{A}, \mathbf{B}) = \prod_{h=1}^H r_{a_h}(\mathbf{a}_h) \prod_{h=1}^H r_{b_h}(\mathbf{b}_h). \quad (3.29)$$

By applying the same procedure as that with the matrixwise independence constraint, we can derive the solution to

$$\widehat{r} = \underset{r}{\operatorname{argmin}} F(r) \quad \text{s.t.} \quad r(\mathbf{A}, \mathbf{B}) = \prod_{h=1}^H r_{a_h}(\mathbf{a}_h) \prod_{h=1}^H r_{b_h}(\mathbf{b}_h), \quad (3.30)$$

which is in the form of the matrix variate Gaussian:

$$\begin{aligned} r_A(\mathbf{A}) &= \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \mathbf{I}_M \otimes \widehat{\Sigma}_A) = \prod_{h=1}^H \text{Gauss}_M(\mathbf{a}_h; \widehat{\mathbf{a}}_h, \widehat{\sigma}_{a_h}^2 \mathbf{I}_M), \\ r_B(\mathbf{B}) &= \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \mathbf{I}_L \otimes \widehat{\Sigma}_B) = \prod_{h=1}^H \text{Gauss}_L(\mathbf{b}_h; \widehat{\mathbf{b}}_h, \widehat{\sigma}_{b_h}^2 \mathbf{I}_L), \end{aligned}$$

with the variational parameters,

$$\begin{aligned} \widehat{\mathbf{A}} &= (\widehat{\mathbf{a}}_1, \dots, \widehat{\mathbf{a}}_H), \\ \widehat{\mathbf{B}} &= (\widehat{\mathbf{b}}_1, \dots, \widehat{\mathbf{b}}_H), \\ \widehat{\Sigma}_A &= \text{Diag}(\widehat{\sigma}_{a_1}^2, \dots, \widehat{\sigma}_{a_H}^2), \\ \widehat{\Sigma}_B &= \text{Diag}(\widehat{\sigma}_{b_1}^2, \dots, \widehat{\sigma}_{b_H}^2). \end{aligned}$$

Here $\text{Diag}(\dots)$ denotes the diagonal matrix with the specified diagonal entries. The stationary conditions are given as follows: for all $h = 1, \dots, H$,

$$\widehat{\mathbf{a}}_h = \frac{\widehat{\sigma}_{a_h}^2}{\sigma^2} \left(\mathbf{V} - \sum_{h' \neq h} \widehat{\mathbf{b}}_{h'} \widehat{\mathbf{a}}_{h'}^\top \right)^\top \widehat{\mathbf{b}}_h, \quad (3.31)$$

$$\widehat{\sigma}_{a_h}^2 = \sigma^2 \left(\left\| \widehat{\mathbf{b}}_h \right\|^2 + L \widehat{\sigma}_{b_h}^2 + \frac{\sigma^2}{c_{a_h}^2} \right)^{-1}, \quad (3.32)$$

$$\widehat{\mathbf{b}}_h = \frac{\widehat{\sigma}_{b_h}^2}{\sigma^2} \left(\mathbf{V} - \sum_{h' \neq h} \widehat{\mathbf{b}}_{h'} \widehat{\mathbf{a}}_{h'}^\top \right) \widehat{\mathbf{a}}_h, \quad (3.33)$$

$$\widehat{\sigma}_{b_h}^2 = \sigma^2 \left(\left\| \widehat{\mathbf{a}}_h \right\|^2 + M \widehat{\sigma}_{a_h}^2 + \frac{\sigma^2}{c_{b_h}^2} \right)^{-1}. \quad (3.34)$$

The free energy is given by Eq. (3.23) with the posterior covariances $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$ restricted to be diagonal. The stationary conditions for the hyperparameters are unchanged, and given by Eqs. (3.26) through (3.28). Therefore, Algorithm 1 with Eqs. (3.31) through (3.34), substituted for Eqs. (3.19) through (3.22), gives a local solution to the VB problem (3.30) with the columnwise independence constraint.

We call this variant *simple VB (SimpleVB) learning*. In Chapter 6, it will be shown that, in the fully observed MF model, the SimpleVB problem (3.30) with columnwise independence and the original VB problem (3.5) with matrixwise independence actually give the equivalent solution.

3.1.2 Special Cases

Probabilistic principal component analysis and reduced rank regression are special cases of matrix factorization. Therefore, they can be trained by Algorithm 1 with or without small modifications.

Probabilistic Principal Component Analysis *Probabilistic principal component analysis* (Tipping and Bishop, 1999; Bishop, 1999b) is a probabilistic model of which the ML estimation corresponds to the classical principal component analysis (PCA) (Hotelling, 1933). The observation $\mathbf{v} \in \mathbb{R}^L$ is assumed to be driven by a latent vector $\widetilde{\mathbf{a}} \in \mathbb{R}^H$ in the following form:

$$\mathbf{v} = \mathbf{B}\widetilde{\mathbf{a}} + \boldsymbol{\varepsilon}.$$

Here, $\mathbf{B} \in \mathbb{R}^{L \times H}$ specifies the linear relationship between $\widetilde{\mathbf{a}}$ and \mathbf{v} , and $\boldsymbol{\varepsilon} \in \mathbb{R}^L$ is a Gaussian noise subject to $\text{Gauss}_L(\mathbf{0}, \sigma^2 \mathbf{I}_L)$.

Suppose that we are given M observed samples $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_M)$ generated from the latent vectors $\mathbf{A}^\top = (\widetilde{\mathbf{a}}_1, \dots, \widetilde{\mathbf{a}}_M)$, and each latent vector is subject to $\widetilde{\mathbf{a}} \sim \text{Gauss}_H(\mathbf{0}, \mathbf{I}_H)$. Then, the probabilistic PCA model is written as Eqs. (3.1) and (3.2) with $\mathbf{C}_A = \mathbf{I}_H$. Having the prior (3.2) on \mathbf{B} , it is equivalent to the MF model.

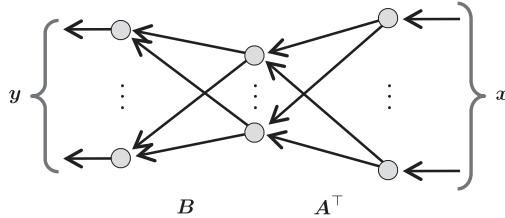


Figure 3.1 Reduced rank regression model.

If we apply VB or EVB learning, the intrinsic dimension H is automatically selected without additional procedure (Bishop, 1999b). This useful property is caused by the ARD (Neal, 1996), which makes the estimators for the irrelevant column vectors of A and B zero. In Chapter 7, this phenomenon is explained in terms of *model-induced regularization (MIR)*, while in Chapter 8, a theoretical guarantee of the dimensionality estimation is given.

Reduced Rank Regression *Reduced rank regression (RRR)* (Baldi and Hornik, 1995; Reinsel and Velu, 1998) is aimed at learning a relation between an input vector $x \in \mathbb{R}^M$ and an output vector $y \in \mathbb{R}^L$ by using the following linear model:

$$y = BA^\top x + \varepsilon, \quad (3.35)$$

where $A \in \mathbb{R}^{M \times H}$ and $B \in \mathbb{R}^{L \times H}$ are parameters to be estimated, and $\varepsilon \sim \text{Gauss}_L(\mathbf{0}, \sigma'^2 I_L)$ is a Gaussian noise. RRR can be seen as a linear neural network (Figure 3.1), of which the model distribution is given by

$$p(y|x, A, B) = (2\pi\sigma'^2)^{-L/2} \exp\left(-\frac{1}{2\sigma'^2} \|y - BA^\top x\|^2\right). \quad (3.36)$$

Thus, we can interpret this model as first projecting the input vector x onto a lower-dimensional latent subspace by A^\top and then performing linear prediction by B .

Suppose we are given N pairs of input and output vectors:

$$\mathcal{D} = \{(x^{(n)}, y^{(n)}) | x^{(n)} \in \mathbb{R}^M, y^{(n)} \in \mathbb{R}^L, n = 1, \dots, N\}. \quad (3.37)$$

Then, the likelihood of the RRR model (3.36) is expressed as

$$\begin{aligned} p(\mathcal{D}|A, B) &= \prod_{n=1}^N p(y^{(n)}|x^{(n)}, A, B)p(x^{(n)}) \\ &\propto \exp\left(-\frac{1}{2\sigma'^2} \sum_{n=1}^N \|y^{(n)} - BA^\top x^{(n)}\|^2\right). \end{aligned} \quad (3.38)$$

Note that we here ignored the input distributions $\prod_{n=1}^N p(\mathbf{x}^{(n)})$ as constants (see Example 1.2 in Section 1.1.1). Let us assume that the samples are *centered*:

$$\frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)} = \mathbf{0} \quad \text{and} \quad \frac{1}{N} \sum_{n=1}^N \mathbf{y}^{(n)} = \mathbf{0}.$$

Furthermore, let us assume that the input samples are *prewhitened* (Hyvärinen et al., 2001), i.e., they satisfy

$$\frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)} \mathbf{x}^{(n)\top} = \mathbf{I}_M.$$

Let

$$\mathbf{V} = \Sigma_{XY} = \frac{1}{N} \sum_{n=1}^N \mathbf{y}^{(n)} \mathbf{x}^{(n)\top} \quad (3.39)$$

be the sample *cross-covariance* matrix, and

$$\sigma^2 = \frac{\sigma'^2}{N} \quad (3.40)$$

be a rescaled noise variance. Then the exponent of the likelihood (3.38) can be written as

$$\begin{aligned} & -\frac{1}{2\sigma'^2} \sum_{n=1}^N \|\mathbf{y}^{(n)} - \mathbf{B}\mathbf{A}^\top \mathbf{x}^{(n)}\|^2 \\ &= -\frac{1}{2\sigma'^2} \sum_{n=1}^N \left\{ \|\mathbf{y}^{(n)}\|^2 - 2\text{tr}(\mathbf{y}^{(n)} \mathbf{x}^{(n)\top} \mathbf{A}\mathbf{B}^\top) + \text{tr}(\mathbf{B}\mathbf{A}^\top \mathbf{x}^{(n)} \mathbf{x}^{(n)\top} \mathbf{A}\mathbf{B}^\top) \right\} \\ &= -\frac{1}{2\sigma'^2} \left\{ \sum_{n=1}^N \|\mathbf{y}^{(n)}\|^2 - 2N\text{tr}(\mathbf{V}\mathbf{A}\mathbf{B}^\top) + N\text{tr}(\mathbf{A}\mathbf{B}^\top \mathbf{B}\mathbf{A}^\top) \right\} \\ &= -\frac{1}{2\sigma'^2} \left(N \|\mathbf{V} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2 + \sum_{n=1}^N \|\mathbf{y}^{(n)}\|^2 - N \|\mathbf{V}\|_{\text{Fro}}^2 \right) \\ &= -\frac{1}{2\sigma^2} \|\mathbf{V} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2 - \frac{1}{2\sigma^2} \left(\frac{1}{N} \sum_{n=1}^N \|\mathbf{y}^{(n)}\|^2 - \|\mathbf{V}\|_{\text{Fro}}^2 \right). \end{aligned} \quad (3.41)$$

The first term in Eq. (3.41) coincides with the log-likelihood of the MF model (3.1), and the second term is constant with respect to \mathbf{A} and \mathbf{B} . Thus, RRR is reduced to MF, as far as the posteriors for \mathbf{A} and \mathbf{B} are concerned.

However, the second term depends on the rescaled noise variance σ^2 , and therefore, should be considered when σ^2 is estimated based on the free energy minimization principle. Furthermore, the normalization constant of the

likelihood (3.38) differs from that of the MF model. Taking these differences into account, the VB free energy of the RRR model (3.38) with the priors (3.2) and (3.3) is given by

$$\begin{aligned} 2F^{\text{RRR}} = & NL \log(2\pi N \sigma^2) + \frac{\frac{1}{N} \sum_{n=1}^N \|y^{(n)}\|^2 - \|\mathbf{V}\|_{\text{Fro}}^2}{\sigma^2} \\ & + \frac{\|\mathbf{V} - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top\|_{\text{Fro}}^2}{\sigma^2} + M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + L \log \frac{\det(\mathbf{C}_B)}{\det(\widehat{\Sigma}_B)} \\ & - (L+M)H + \text{tr} \left\{ \mathbf{C}_A^{-1} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) + \mathbf{C}_B^{-1} \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B \right) \right. \\ & \left. + \sigma^{-2} \left(-\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + (\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A)(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B) \right) \right\}. \end{aligned} \quad (3.42)$$

Note that the difference from Eq. (3.23) is only in the first two terms. Accordingly, the stationary conditions for the variational parameters $\widehat{\mathbf{A}}$, $\widehat{\mathbf{B}}$, $\widehat{\Sigma}_A$, and $\widehat{\Sigma}_B$, and those for the prior covariances \mathbf{C}_A and \mathbf{C}_B (in EVB learning) are the same, i.e., the update rules given by Eqs. (3.19) through (3.22), (3.26), and (3.27) are valid for the RRR model. The update rule for the rescaled noise variance is different from Eq. (3.28), and given by

$$(\sigma^2)^{\text{RRR}} = \frac{\frac{1}{N} \sum_{n=1}^N \|y^{(n)}\|^2 - \text{tr} \left(2\mathbf{V}^\top \widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top \right) + \text{tr} \left((\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A)(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B) \right)}{NL}, \quad (3.43)$$

which was obtained from the derivative of Eq. (3.42), instead of Eq. (3.23), with respect to σ^2 .

Once the rescaled noise variance σ^2 is estimated, Eq. (3.40) gives the original noise variance σ'^2 of the RRR model (3.38).

3.2 Matrix Factorization with Missing Entries

One of the major applications of MF is *collaborative filtering (CF)*, where only a part of the entries in \mathbf{V} are observed, and the task is to predict missing entries. We can derive a VB algorithm for this scenario, similarly to the fully observed case.

3.2.1 VB Learning for MF with Missing Entries

To express missing entries, the likelihood (3.1) should be replaced with

$$p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \propto \exp \left(-\frac{1}{2\sigma^2} \left\| \mathcal{P}_A(\mathbf{V}) - \mathcal{P}_A(\mathbf{B}\mathbf{A}^\top) \right\|_{\text{Fro}}^2 \right), \quad (3.44)$$

where Λ denotes the set of observed indices, and $\mathcal{P}_\Lambda(\mathbf{V})$ denotes the matrix of the same size as \mathbf{V} with its entries given by

$$(\mathcal{P}_\Lambda(\mathbf{V}))_{l,m} = \begin{cases} V_{l,m} & \text{if } (l, m) \in \Lambda, \\ 0 & \text{otherwise.} \end{cases}$$

Conditional Conjugacy

Since the likelihood (3.44) is still in a Gaussian form of \mathbf{A} if \mathbf{B} is regarded as a constant, or vice versa, the conditional conjugacy with respect to \mathbf{A} and \mathbf{B} , respectively, still holds if we adopt the Gaussian priors (3.2) and (3.3):

$$\begin{aligned} p(\mathbf{A}) &\propto \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{A}\mathbf{C}_A^{-1}\mathbf{A}^\top\right)\right), \\ p(\mathbf{B}) &\propto \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{B}\mathbf{C}_B^{-1}\mathbf{B}^\top\right)\right). \end{aligned}$$

The posterior will be still Gaussian, but in a broader class than the fully observed case, as will be seen shortly.

Variational Bayesian Algorithm

With the missing entries, the stationary condition (3.7) becomes

$$\begin{aligned} r_A(\mathbf{A}) &\propto \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{A}\mathbf{C}_A^{-1}\mathbf{A}^\top\right) - \frac{1}{2\sigma^2} \left\langle \left\| \mathcal{P}_\Lambda(\mathbf{V}) - \mathcal{P}_\Lambda(\mathbf{B}\mathbf{A}^\top) \right\|^2_{\text{Fro}} \right\rangle_{r_B(\mathbf{B})}\right) \\ &\propto \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{A}\mathbf{C}_A^{-1}\mathbf{A}^\top\right) + \sigma^{-2} \sum_{(l,m) \in \Lambda} \left\langle -2V_{l,m} \sum_{h=1}^H B_{l,h} A_{m,h} + \sum_{h=1}^H \sum_{h'=1}^H B_{l,h} B_{l,h'} A_{m,h} A_{m,h'} \right\rangle_{r_B(\mathbf{B})}\right) \\ &\propto \exp\left(-\frac{\sum_{m=1}^M \left\langle (\tilde{\mathbf{a}}_m - \widehat{\tilde{\Sigma}}_{A,m}^{-1}(\tilde{\mathbf{a}}_m - \widehat{\tilde{\Sigma}}_{A,m}^{-1}\tilde{\mathbf{a}}_m))^\top \widehat{\tilde{\Sigma}}_{A,m}^{-1}(\tilde{\mathbf{a}}_m - \widehat{\tilde{\Sigma}}_{A,m}^{-1}\tilde{\mathbf{a}}_m) \right\rangle}{2}\right), \end{aligned} \quad (3.45)$$

where

$$\widehat{\tilde{\mathbf{a}}}_m = \sigma^{-2} \widehat{\tilde{\Sigma}}_{A,m} \sum_{l:(l,m) \in \Lambda} V_{l,m} \left\langle \tilde{\mathbf{b}}_l \right\rangle_{r_B(\mathbf{B})}, \quad (3.46)$$

$$\widehat{\tilde{\Sigma}}_{A,m} = \sigma^2 \left(\sum_{l:(l,m) \in \Lambda} \left\langle \tilde{\mathbf{b}}_l \tilde{\mathbf{b}}_l^\top \right\rangle_{r_B(\mathbf{B})} + \sigma^2 \mathbf{C}_A^{-1} \right)^{-1}. \quad (3.47)$$

Here, $\sum_{(l,m) \in \Lambda}$ denotes the sum over l and m such that $(l, m) \in \Lambda$, and $\sum_{l:(l,m) \in \Lambda}$ denotes the sum over l such that $(l, m) \in \Lambda$ for given m .

Similarly, we have

$$\begin{aligned} r_B(\mathbf{B}) &\propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{B}\mathbf{C}_B^{-1}\mathbf{B}^\top) - \frac{1}{2\sigma^2} \left\langle \left\| \mathcal{P}_A(V) - \mathcal{P}_A(\mathbf{B}\mathbf{A}^\top) \right\|_{\text{Fro}}^2 \right\rangle_{r_A(A)}\right) \\ &\propto \exp\left(-\frac{\sum_{l=1}^L (\tilde{\mathbf{b}}_m - \widehat{\tilde{\mathbf{b}}}_l)^\top \widehat{\Sigma}_{B,l}^{-1} (\tilde{\mathbf{b}}_l - \widehat{\tilde{\mathbf{b}}}_l)}{2}\right), \end{aligned} \quad (3.48)$$

where

$$\widehat{\tilde{\mathbf{b}}}_l = \sigma^{-2} \widehat{\Sigma}_{B,l} \sum_{m:(l,m) \in A} V_{l,m} \langle \tilde{\mathbf{a}}_m \rangle_{r_A(A)}, \quad (3.49)$$

$$\widehat{\Sigma}_{B,l} = \sigma^2 \left(\sum_{m:(l,m) \in A} \langle \tilde{\mathbf{a}}_m \tilde{\mathbf{a}}_m^\top \rangle_{r_A(A)} + \sigma^2 \mathbf{C}_B^{-1} \right)^{-1}. \quad (3.50)$$

Eqs. (3.45) and (3.48) imply that \mathbf{A} and \mathbf{B} are Gaussian in the following form:

$$\begin{aligned} r_A(\mathbf{A}) &= \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \check{\Sigma}_A) = \prod_{m=1}^M \text{Gauss}_H(\tilde{\mathbf{a}}_m; \widehat{\tilde{\mathbf{a}}}_m, \widehat{\Sigma}_{A,m}), \\ r_B(\mathbf{B}) &= \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \check{\Sigma}_B) = \prod_{l=1}^L \text{Gauss}_H(\tilde{\mathbf{b}}_l; \widehat{\tilde{\mathbf{b}}}_l, \widehat{\Sigma}_{B,l}), \end{aligned}$$

where

$$\check{\Sigma}_A = \begin{pmatrix} \widehat{\Sigma}_{A,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\Sigma}_{A,2} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \widehat{\Sigma}_{A,M} \end{pmatrix}, \quad \check{\Sigma}_B = \begin{pmatrix} \widehat{\Sigma}_{B,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\Sigma}_{B,2} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \widehat{\Sigma}_{B,L} \end{pmatrix}.$$

Note that the posterior covariances cannot be expressed with a Kronecker product, unlike the fully observed case. However, the posteriors are Gaussian, and moments are given by

$$\begin{aligned} \langle \tilde{\mathbf{a}}_m \rangle_{r_A(A)} &= \widehat{\tilde{\mathbf{a}}}_m, \\ \langle \tilde{\mathbf{a}}_m \tilde{\mathbf{a}}_m^\top \rangle_{r_A(A)} &= \widehat{\tilde{\mathbf{a}}}_m \widehat{\tilde{\mathbf{a}}}_m^\top + \widehat{\Sigma}_{A,m}, \\ \langle \tilde{\mathbf{b}}_l \rangle_{r_B(B)} &= \widehat{\tilde{\mathbf{b}}}_l, \\ \langle \tilde{\mathbf{b}}_l \tilde{\mathbf{b}}_l^\top \rangle_{r_B(B)} &= \widehat{\tilde{\mathbf{b}}}_l \widehat{\tilde{\mathbf{b}}}_l^\top + \widehat{\Sigma}_{B,l}. \end{aligned}$$

Thus, Eqs. (3.46), (3.47), (3.49), and (3.50) lead to

$$\widehat{\bar{\mathbf{a}}}_m = \sigma^{-2} \widehat{\Sigma}_{A,m} \sum_{l:(l,m) \in \Lambda} V_{l,m} \widehat{\bar{\mathbf{b}}}_l, \quad (3.51)$$

$$\widehat{\Sigma}_{A,m} = \sigma^2 \left(\sum_{l:(l,m) \in \Lambda} \left(\widehat{\bar{\mathbf{b}}}_l \widehat{\bar{\mathbf{b}}}_l^\top + \widehat{\Sigma}_{B,l} \right) + \sigma^2 \mathbf{C}_A^{-1} \right)^{-1}, \quad (3.52)$$

$$\widehat{\bar{\mathbf{b}}}_l = \sigma^{-2} \widehat{\Sigma}_{B,l} \sum_{m:(l,m) \in \Lambda} V_{l,m} \widehat{\bar{\mathbf{a}}}_m, \quad (3.53)$$

$$\widehat{\Sigma}_{B,l} = \sigma^2 \left(\sum_{m:(l,m) \in \Lambda} \left(\widehat{\bar{\mathbf{a}}}_m \widehat{\bar{\mathbf{a}}}_m^\top + \widehat{\Sigma}_{A,m} \right) + \sigma^2 \mathbf{C}_B^{-1} \right)^{-1}, \quad (3.54)$$

which are used as update rules for local search (Lim and Teh, 2007).

Free Energy as a Function of Variational Parameters

An explicit form of the free energy can be obtained in a similar fashion to the fully observed case:

$$\begin{aligned} 2F = & \#(\Lambda) \cdot \log(2\pi\sigma^2) + M \log \det(\mathbf{C}_A) + L \log \det(\mathbf{C}_B) \\ & - \sum_{m=1}^M \log \det(\widehat{\Sigma}_{A,m}) - \sum_{l=1}^L \log \det(\widehat{\Sigma}_{B,l}) - (L+M)H \\ & + \text{tr} \left\{ \mathbf{C}_A^{-1} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + \sum_{m=1}^M \widehat{\Sigma}_{A,m} \right) + \mathbf{C}_B^{-1} \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + \sum_{l=1}^L \widehat{\Sigma}_{B,l} \right) \right\} \\ & + \sigma^{-2} \sum_{(l,m) \in \Lambda} \left(V_{l,m} - 2V_{l,m} \widehat{\bar{\mathbf{a}}}_m^\top \widehat{\bar{\mathbf{b}}}_l + \text{tr} \left\{ \left(\widehat{\bar{\mathbf{a}}}_m \widehat{\bar{\mathbf{a}}}_m^\top + \widehat{\Sigma}_{A,m} \right) \left(\widehat{\bar{\mathbf{b}}}_l \widehat{\bar{\mathbf{b}}}_l^\top + \widehat{\Sigma}_{B,l} \right) \right\} \right), \end{aligned} \quad (3.55)$$

where $\#(\Lambda)$ denotes the number of observed entries.

Empirical Variational Bayesian Algorithm

By taking derivatives of the free energy (3.55), we can derive update rules for the hyperparameters:

$$c_{a_h}^2 = \frac{\|\widehat{\bar{\mathbf{a}}}_h\|^2 + (\sum_{m=1}^M \widehat{\Sigma}_{A,m})_{h,h}}{M}, \quad (3.56)$$

$$c_{b_h}^2 = \frac{\|\widehat{\bar{\mathbf{b}}}_h\|^2 + (\sum_{l=1}^L \widehat{\Sigma}_{B,l})_{h,h}}{L}, \quad (3.57)$$

Algorithm 2 EVB learning for matrix factorization with missing entries.

-
- 1: Initialize the variational parameters $(\widehat{\mathbf{A}}, \{\widehat{\Sigma}_{A,m}\}_{m=1}^M, \widehat{\mathbf{B}}, \{\widehat{\Sigma}_{B,l}\}_{l=1}^L)$, and the hyperparameters $(\mathbf{C}_A, \mathbf{C}_B, \sigma^2)$, for example, $\widehat{\mathbf{A}}_{m,h}, \widehat{\mathbf{B}}_{l,h} \sim \text{Gauss}_1(0, \tau)$, $\widehat{\Sigma}_{A,m} = \widehat{\Sigma}_{B,l} = \mathbf{C}_A = \mathbf{C}_B = \tau \mathbf{I}_H$, and $\sigma^2 = \tau^2 = \sum_{(l,m) \in \Lambda} V_{l,m}^2 / \#(\Lambda)$.
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (3.52), (3.51), (3.54), and (3.53) to update $\{\widehat{\Sigma}_{A,m}\}_{m=1}^M$, $\widehat{\mathbf{A}}$, $\{\widehat{\Sigma}_{B,l}\}_{l=1}^L$, and $\widehat{\mathbf{B}}$, respectively.
 - 3: Apply Eqs. (3.56) and (3.57) for all $h = 1, \dots, H$, and Eq. (3.58) to update \mathbf{C}_A , \mathbf{C}_B , and σ^2 , respectively.
 - 4: Prune the h th component if $c_{a_h}^2 c_{b_h}^2 < \varepsilon$, where $\varepsilon > 0$ is a threshold, e.g., set to $\varepsilon = 10^{-4}$.
 - 5: Evaluate the free energy (3.55).
 - 6: Iterate Steps 2 through 5 until convergence (until the energy decrease becomes smaller than a threshold).
-

$$\sigma^2 = \frac{\sum_{(l,m) \in \Lambda} \left(V_{l,m} - 2V_{l,m} \widehat{\mathbf{a}}_m^\top \widehat{\mathbf{b}}_l + \text{tr} \left\{ \left(\widehat{\mathbf{a}}_m \widehat{\mathbf{a}}_m^\top + \widehat{\Sigma}_{A,m} \right) \left(\widehat{\mathbf{b}}_l \widehat{\mathbf{b}}_l^\top + \widehat{\Sigma}_{B,l} \right) \right\} \right)}{\#(\Lambda)}. \quad (3.58)$$

Algorithm 2 summarizes the EVB algorithm for MF with missing entries. Again, if we appropriately set the hyperparameters $(\mathbf{C}_A, \mathbf{C}_B, \sigma^2)$ in Step 1 and skip Steps 3 and 4, Algorithm 2 is reduced to (nonempirical) VB learning.

Simple Variational Bayesian Learning (with Columnwise Independence)
Similarly to the fully observed case, we can reduce the computational burden and the memory requirement of VB learning by adopting the columnwise independence (Ilin and Raiko, 2010):

$$r(\mathbf{A}, \mathbf{B}) = \prod_{h=1}^H r_{a_h}(\mathbf{a}_h) \prod_{h=1}^H r_{b_h}(\mathbf{b}_h). \quad (3.59)$$

By applying the same procedure as the matrixwise independence case, we can derive the solution to

$$\widehat{r} = \underset{r}{\operatorname{argmin}} F(r) \quad \text{s.t.} \quad r(\mathbf{A}, \mathbf{B}) = \prod_{h=1}^H r_{a_h}(\mathbf{a}_h) \prod_{h=1}^H r_{b_h}(\mathbf{b}_h), \quad (3.60)$$

which is in the form of the matrix variate Gaussian,

$$r_A(\mathbf{A}) = \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \check{\Sigma}_A) = \prod_{m=1}^M \prod_{h=1}^H \text{Gauss}_1(A_{m,h}; \widehat{A}_{m,h}, \widehat{\sigma}_{A_{m,h}}^2),$$

$$r_B(\mathbf{B}) = \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \check{\Sigma}_B) = \prod_{l=1}^L \prod_{h=1}^H \text{Gauss}_1(B_{l,h}; \widehat{B}_{l,h}, \widehat{\sigma}_{B_{l,h}}^2),$$

with diagonal posterior covariances, i.e.,

$$\check{\Sigma}_A = \begin{pmatrix} \widehat{\Sigma}_{A,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\Sigma}_{A,2} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \widehat{\Sigma}_{A,M} \end{pmatrix}, \quad \check{\Sigma}_B = \begin{pmatrix} \widehat{\Sigma}_{B,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\Sigma}_{B,2} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \widehat{\Sigma}_{B,L} \end{pmatrix},$$

for

$$\widehat{\Sigma}_{A,m} = \text{Diag}(\widehat{\sigma}_{A_{m,1}}^2, \dots, \widehat{\sigma}_{A_{m,H}}^2),$$

$$\widehat{\Sigma}_{B,l} = \text{Diag}(\widehat{\sigma}_{B_{l,1}}^2, \dots, \widehat{\sigma}_{B_{l,H}}^2).$$

The stationary conditions are given as follows: for all $l = 1, \dots, L$, $m = 1, \dots, M$, and $h = 1, \dots, H$,

$$\widehat{A}_{m,h} = \frac{\widehat{\sigma}_{A_{m,h}}^2}{\sigma^2} \sum_{l:(l,m) \in \Lambda} \left(V_{l,m} - \sum_{h' \neq h} \widehat{B}_{l,h'} \widehat{A}_{m,h'} \right) \widehat{B}_{l,h}, \quad (3.61)$$

$$\widehat{\sigma}_{A_{m,h}}^2 = \sigma^2 \left(\sum_{l:(l,m) \in \Lambda} \left(\widehat{B}_{l,h}^2 + \widehat{\sigma}_{B_{l,h}}^2 \right) + \frac{\sigma^2}{c_{a_h}^2} \right)^{-1}, \quad (3.62)$$

$$\widehat{B}_{l,h} = \frac{\widehat{\sigma}_{B_{l,h}}^2}{\sigma^2} \sum_{m:(l,m) \in \Lambda} \left(V_{l,m} - \sum_{h' \neq h} \widehat{A}_{m,h'} \widehat{B}_{l,h'} \right) \widehat{A}_{m,h}, \quad (3.63)$$

$$\widehat{\sigma}_{B_{l,h}}^2 = \sigma^2 \left(\sum_{m:(l,m) \in \Lambda} \left(\widehat{A}_{m,h}^2 + \widehat{\sigma}_{A_{m,h}}^2 \right) + \frac{\sigma^2}{c_{b_h}^2} \right)^{-1}. \quad (3.64)$$

The free energy is given by Eq. (3.55) with the posterior covariances $\{\widehat{\Sigma}_{A,m}, \widehat{\Sigma}_{B,l}\}$ restricted to be diagonal. The stationary conditions for the hyperparameters are unchanged, and given by Eqs. (3.56) through (3.58). Therefore, Algorithm 2 with Eqs. (3.61) through (3.64) substituted for Eqs. (3.51) through (3.54) gives a local solution to the VB problem (3.60) with the columnwise independence constraint.

SimpleVB learning is much more practical when missing entries exist. In the fully observed case, the posterior covariances $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$ are common to all rows of \mathbf{A} and to all rows of \mathbf{B} , respectively, while in the partially

observed case, we need to store the posterior covariances $\widehat{\Sigma}_{A,m}$ and $\widehat{\Sigma}_{B,l}$ for all $m = 1, \dots, M$ and $l = 1, \dots, L$. Since L and M can be huge, e.g., in collaborative filtering applications, the required memory size is significantly reduced by restricting the covariances to be diagonal.

3.3 Tensor Factorization

A matrix is a two-dimensional array of numbers. We can extend this notion to an N -dimensional array, which is called an N -mode tensor. Namely, a tensor $\mathcal{V} \in \mathbb{R}^{M^{(1)} \times \dots \times M^{(N)}}$ consists of $\prod_{n=1}^N M^{(n)}$ entries lying in an N -dimensional array, where $M^{(n)}$ denotes the length in mode n . In this section, we derive VB learning for tensor factorization.

3.3.1 Tucker Factorization

Similarly to the rank of a matrix, we can control the degree of freedom of a tensor by controlling its *tensor rank*. Although there are a few different definitions of the tensor rank and corresponding ways of factorization, we here focus on *Tucker factorization (TF)* (Tucker, 1996; Kolda and Bader, 2009) defined as follows:

$$\mathcal{V} = \mathcal{G} \times_1 \mathbf{A}^{(1)} \cdots \times_N \mathbf{A}^{(N)} + \mathcal{E},$$

where $\mathcal{V} \in \mathbb{R}^{M^{(1)} \times \dots \times M^{(N)}}$, $\mathcal{G} \in \mathbb{R}^{H^{(1)} \times \dots \times H^{(N)}}$, and $\{\mathbf{A}^{(n)} \in \mathbb{R}^{M^{(n)} \times H^{(n)}}\}$ are an observed tensor, a *core tensor*, and *factor matrices*, respectively. $\mathcal{E} \in \mathbb{R}^{M^{(1)} \times \dots \times M^{(N)}}$ is noise and \times_n denotes the n -mode tensor product. *Parafac* (Harshman, 1970), another popular way of tensor factorization, can be seen as a special case of Tucker factorization where the core tensor is *superdiagonal*, i.e., only the entries $G_{h^{(1)}, \dots, h^{(N)}}$ for $h^{(1)} = h^{(2)} = \dots = h^{(N)}$ are nonzero.

3.3.2 VB Learning for TF

The probabilistic model for MF is straightforwardly extended to TF (Chu and Ghahramani, 2009; Mørup and Hansen, 2009). Assume Gaussian noise and Gaussian priors:

$$p(\mathcal{V}|\mathcal{G}, \{\mathbf{A}^{(n)}\}) \propto \exp\left(-\frac{\|\mathcal{V} - \mathcal{G} \times_1 \mathbf{A}^{(1)} \cdots \times_N \mathbf{A}^{(N)}\|^2}{2\sigma^2}\right), \quad (3.65)$$

$$p(\mathcal{G}) \propto \exp\left(-\frac{\text{vec}(\mathcal{G})^\top (\mathbf{C}_{G^{(N)}} \otimes \cdots \otimes \mathbf{C}_{G^{(1)}})^{-1} \text{vec}(\mathcal{G})}{2}\right), \quad (3.66)$$

$$p(\{\mathbf{A}^{(n)}\}) \propto \exp\left(-\frac{\sum_{n=1}^N \text{tr}(\mathbf{A}^{(n)} \mathbf{C}_{A^{(n)}}^{-1} \mathbf{A}^{(n)\top})}{2}\right), \quad (3.67)$$

where \otimes and $\text{vec}(\cdot)$ denote the *Kronecker product* and the *vectorization operator*, respectively. $\{\mathbf{C}_{G^{(n)}}\}$ and $\{\mathbf{C}_{A^{(n)}}\}$ are the prior covariances restricted to be diagonal, i.e.,

$$\begin{aligned} \mathbf{C}_{G^{(n)}} &= \mathbf{Diag}\left(c_{g_1^{(n)}}^2, \dots, c_{g_{H^{(n)}}^{(n)}}^2\right), \\ \mathbf{C}_{A^{(n)}} &= \mathbf{Diag}\left(c_{a_1^{(n)}}^2, \dots, c_{a_{H^{(n)}}^{(n)}}^2\right). \end{aligned}$$

We denote $\check{\mathbf{C}}_G = \mathbf{C}_{G^{(N)}} \otimes \cdots \otimes \mathbf{C}_{G^{(1)}}$.

Conditional Conjugacy

Since the TF model is multilinear, the likelihood (3.65) is in the Gaussian form of the core tensor \mathcal{G} and of each of the factor matrices $\{\mathbf{A}^{(n)}\}$, if the others are fixed as constants. Therefore, the Gaussian priors (3.66) and (3.67) are conditionally conjugate for each parameter, and the posterior will be Gaussian.

Variational Bayesian Algorithm

Based on the conditional conjugacy, we impose the following constraint on the VB posterior:

$$r(\mathcal{G}, \{\mathbf{A}^{(n)}\}) = r(\mathcal{G}) \prod_{n=1}^N r(\mathbf{A}^{(n)}).$$

Then, the free energy can be written as

$$\begin{aligned} F(r) &= \int r(\mathcal{G}) \left(\prod_{n=1}^N r(\mathbf{A}^{(n)}) \right) \left(\log p(\mathcal{V}, \mathcal{G}, \{\mathbf{A}^{(n)}\}) - \log r(\mathcal{G}) - \sum_{n=1}^N \log r(\mathbf{A}^{(n)}) \right) \\ &\quad \cdot d\mathcal{G} \left(\prod_{n=1}^N d\mathbf{A}^{(n)} \right). \end{aligned} \quad (3.68)$$

Using the variational method, we obtain

$$\begin{aligned} 0 &= \int \left(\prod_{n=1}^N r(\mathbf{A}^{(n)}) \right) \left(\log p(\mathcal{V}, \mathcal{G}, \{\mathbf{A}^{(n)}\}) - \log r(\mathcal{G}) - \sum_{n=1}^N \log r(\mathbf{A}^{(n)}) - 1 \right) \\ &\quad \cdot \left(\prod_{n=1}^N d\mathbf{A}^{(n)} \right), \end{aligned}$$

and therefore

$$\begin{aligned} r(\mathcal{G}) &\propto \exp\langle\log p(\mathcal{V}, \mathcal{G}, \{\mathbf{A}^{(n)}\})\rangle_{r(\{\mathbf{A}^{(n)}\})} \\ &\propto p(\mathcal{G}) \exp\langle\log p(\mathcal{V}|\mathcal{G}, \{\mathbf{A}^{(n)}\})\rangle_{r(\{\mathbf{A}^{(n)}\})}. \end{aligned} \quad (3.69)$$

Similarly, we can also obtain

$$\begin{aligned} 0 = \int r(\mathcal{G}) \left(\prod_{n' \neq n} r(\mathbf{A}^{(n')}) \right) &\left(\log p(\mathcal{V}, \mathcal{G}, \{\mathbf{A}^{(n)}\}) - \log r(\mathcal{G}) - \sum_{n=1}^N \log r(\mathbf{A}^{(n)}) - 1 \right) \\ &\cdot \left(\prod_{n' \neq n} d\mathbf{A}^{(n')} \right), \end{aligned}$$

and therefore

$$\begin{aligned} r(\mathbf{A}^{(n)}) &\propto \exp\langle\log p(\mathcal{V}, \mathcal{G}, \{\mathbf{A}^{(n)}\})\rangle_{r(\mathcal{G})r(\{\mathbf{A}^{(n')}\}_{n' \neq n})} \\ &\propto p(\mathbf{A}^{(n)}) \exp\langle\log p(\mathcal{V}|\mathcal{G}, \{\mathbf{A}^{(n)}\})\rangle_{r(\mathcal{G})r(\{\mathbf{A}^{(n')}\}_{n' \neq n})}. \end{aligned} \quad (3.70)$$

Eqs. (3.69) and (3.70) imply that the VB posteriors are Gaussian. The expectation in Eq. (3.69) can be calculated as follows:

$$\begin{aligned} \langle\log p(\mathcal{V}|\mathcal{G}, \{\mathbf{A}^{(n)}\})\rangle_{r(\{\mathbf{A}^{(n)}\})} &= -\frac{1}{2\sigma^2} \left\langle \left\| \mathcal{V} - \mathcal{G} (\times_1 \mathbf{A}^{(1)} \otimes \cdots \otimes \times_N \mathbf{A}^{(N)}) \right\|^2 \right\rangle_{r(\{\mathbf{A}^{(n)}\})} + \text{const.} \\ &= -\frac{1}{2\sigma^2} \left\langle -2\text{vec}(\mathcal{V})^\top (\mathbf{A}^{(N)} \otimes \cdots \otimes \mathbf{A}^{(1)}) \text{vec}(\mathcal{G}) \right. \\ &\quad \left. + \text{vec}(\mathcal{G})^\top (\mathbf{A}^{(N)\top} \mathbf{A}^{(N)} \otimes \cdots \otimes \mathbf{A}^{(1)\top} \mathbf{A}^{(1)}) \text{vec}(\mathcal{G}) \right\rangle_{r(\{\mathbf{A}^{(n)}\})} + \text{const.} \end{aligned}$$

Substituting the preceding calculation and the prior (3.66) into Eq. (3.69) gives

$$\begin{aligned} \log r(\mathcal{G}) &= \log p(\mathcal{G}) \langle\log p(\mathcal{V}|\mathcal{G}, \{\mathbf{A}^{(n)}\})\rangle_{r(\{\mathbf{A}^{(n)}\})} + \text{const.} \\ &= -\frac{1}{2} (\check{\mathbf{g}} - \widehat{\check{\mathbf{g}}})^\top \widetilde{\Sigma}_G^{-1} (\check{\mathbf{g}} - \widehat{\check{\mathbf{g}}}) + \text{const.}, \end{aligned}$$

where

$$\begin{aligned} \check{\mathbf{g}} &= \text{vec}(\mathcal{G}), \\ \check{\mathbf{v}} &= \text{vec}(\mathcal{V}), \\ \widehat{\check{\mathbf{g}}} &= \frac{\widetilde{\Sigma}_G}{\sigma^2} \left(\widehat{\mathbf{A}}^{(N)} \otimes \cdots \otimes \widehat{\mathbf{A}}^{(1)} \right)^\top \check{\mathbf{v}}, \end{aligned} \quad (3.71)$$

$$\widetilde{\Sigma}_G = \sigma^2 \left(\left\langle \mathbf{A}^{(N)\top} \mathbf{A}^{(N)} \otimes \cdots \otimes \mathbf{A}^{(1)\top} \mathbf{A}^{(1)} \right\rangle_{r(\{\mathbf{A}^{(n)}\})} + \sigma^2 \check{\mathbf{C}}_G^{-1} \right)^{-1}. \quad (3.72)$$

Similarly, the expectation in Eq. (3.70) can be calculated as follows:

$$\begin{aligned}
& \langle \log p(\mathcal{V}|\mathcal{G}, \mathbf{A}^{(n)}) \rangle_{r(\mathcal{G})r(\{\mathbf{A}^{(n')}\}_{n' \neq n})} \\
&= -\frac{1}{2\sigma^2} \left\langle \left\| \mathcal{V} - \mathcal{G} (\times_1 \mathbf{A}^{(1)}) \cdots (\times_N \mathbf{A}^{(N)}) \right\|^2 \right\rangle_{r(\mathcal{G})r(\{\mathbf{A}^{(n')}\}_{n' \neq n})} + \text{const.} \\
&= -\frac{1}{2\sigma^2} \left\{ \text{tr} \left(-2 \mathbf{V}_{(n)}^\top \mathbf{A}^{(n)} \widehat{\mathbf{G}}_{(n)} (\widehat{\mathbf{A}}^{(N)} \otimes \cdots \otimes \widehat{\mathbf{A}}^{(n+1)} \otimes \widehat{\mathbf{A}}^{(n-1)} \cdots \otimes \widehat{\mathbf{A}}^{(1)})^\top \right) \right. \\
&\quad \left. + \text{tr} \left(\mathbf{A}^{(n)} \left\langle \mathbf{G}_{(n)} (\mathbf{A}^{(N)\top} \mathbf{A}^{(N)} \otimes \cdots \otimes \mathbf{A}^{(n+1)\top} \mathbf{A}^{(n+1)} \right. \right. \right. \\
&\quad \left. \left. \otimes \mathbf{A}^{(n-1)\top} \mathbf{A}^{(n-1)} \cdots \otimes \mathbf{A}^{(1)\top} \mathbf{A}^{(1)}) \mathbf{G}_{(n)}^\top \right\rangle_{r(\mathcal{G})r(\{\mathbf{A}^{(n')}\}_{n' \neq n})} \mathbf{A}^{(n)\top} \right) \right\}.
\end{aligned}$$

Substituting the preceding calculation and the prior (3.67) into Eq. (3.70) gives

$$\begin{aligned}
\log r(\mathbf{A}^{(n)}) &= \log p(\mathbf{A}^{(n)}) \exp \langle \log p(\mathcal{V}|\mathcal{G}, \{\mathbf{A}^{(n')}\}) \rangle_{r(\mathcal{G})r(\{\mathbf{A}^{(n')}\}_{n' \neq n})} + \text{const.} \\
&= -\frac{1}{2} \text{tr} \left((\mathbf{A}^{(n)} - \widehat{\mathbf{A}}^{(n)}) \widehat{\Sigma}_{\mathbf{A}^{(n)}}^{-1} (\mathbf{A}^{(n)} - \widehat{\mathbf{A}}^{(n)})^\top \right),
\end{aligned}$$

where

$$\widehat{\mathbf{A}}^{(n)} = \frac{1}{\sigma^2} \mathbf{V}_{(n)} (\widehat{\mathbf{A}}^{(N)} \otimes \cdots \otimes \widehat{\mathbf{A}}^{(n+1)} \otimes \widehat{\mathbf{A}}^{(n-1)} \cdots \otimes \widehat{\mathbf{A}}^{(1)}) \widehat{\mathbf{G}}_{(n)}^\top \widehat{\Sigma}_{\mathbf{A}^{(n)}}, \quad (3.73)$$

$$\begin{aligned}
\widehat{\Sigma}_{\mathbf{A}^{(n)}} &= \sigma^2 \left(\left\langle \mathbf{G}_{(n)} (\mathbf{A}^{(N)\top} \mathbf{A}^{(N)} \otimes \cdots \otimes \mathbf{A}^{(n+1)\top} \mathbf{A}^{(n+1)} \right. \right. \\
&\quad \left. \left. \otimes \mathbf{A}^{(n-1)\top} \mathbf{A}^{(n-1)} \otimes \cdots \otimes \mathbf{A}^{(1)\top} \mathbf{A}^{(1)}) \mathbf{G}_{(n)}^\top \right\rangle_{r(\mathcal{G})r(\{\mathbf{A}^{(n')}\}_{n' \neq n})} + \sigma^2 \mathbf{C}_{\mathbf{A}^{(n)}}^{-1} \right)^{-1}.
\end{aligned} \quad (3.74)$$

Thus, the VB posterior is given by

$$\begin{aligned}
r(\mathcal{G}, \{\mathbf{A}^{(n)}\}) &= \text{Gauss}_{\prod_{n=1}^N H^{(n)}} \left(\check{\mathbf{g}}; \widehat{\check{\mathbf{g}}}, \widehat{\check{\Sigma}}_G \right) \\
&\quad \cdot \prod_{n=1}^N \text{MGauss}_{M^{(n)}, H^{(n)}} \left(\mathbf{A}^{(n)}; \widehat{\mathbf{A}}^{(n)}, \mathbf{I}_{M^{(n)}} \otimes \widehat{\Sigma}_{\mathbf{A}^{(n)}} \right), \quad (3.75)
\end{aligned}$$

where the means and the covariances satisfy

$$\widehat{\check{\mathbf{g}}} = \frac{\widehat{\check{\Sigma}}_G}{\sigma^2} \left(\widehat{\mathbf{A}}^{(N)} \otimes \cdots \otimes \widehat{\mathbf{A}}^{(1)} \right)^\top \check{\mathbf{v}}, \quad (3.76)$$

$$\begin{aligned}
\widehat{\check{\Sigma}}_G &= \sigma^2 \left(\left(\widehat{\mathbf{A}}^{(N)\top} \widehat{\mathbf{A}}^{(N)} + M^{(N)} \widehat{\Sigma}_{\mathbf{A}^{(N)}} \right) \otimes \cdots \otimes \left(\widehat{\mathbf{A}}^{(1)\top} \widehat{\mathbf{A}}^{(1)} + M^{(1)} \Sigma_{\mathbf{A}^{(1)}} \right) \right. \\
&\quad \left. + \sigma^2 \check{\mathbf{C}}_G^{-1} \right)^{-1}, \quad (3.77)
\end{aligned}$$

$$\widehat{\mathbf{A}}^{(n)} = \frac{1}{\sigma^2} \mathbf{V}_{(n)} (\widehat{\mathbf{A}}^{(N)} \otimes \cdots \otimes \widehat{\mathbf{A}}^{(n+1)} \otimes \widehat{\mathbf{A}}^{(n-1)} \cdots \otimes \widehat{\mathbf{A}}^{(1)}) \widehat{\mathbf{G}}_{(n)}^\top \widehat{\Sigma}_{\mathbf{A}^{(n)}}, \quad (3.78)$$

$$\begin{aligned} \widehat{\Sigma}_{\mathbf{A}^{(n)}} &= \sigma^2 \left(\left\langle \mathbf{G}_{(n)} \left((\widehat{\mathbf{A}}^{(N)\top} \widehat{\mathbf{A}}^{(N)} + M^{(N)} \widehat{\Sigma}_{\mathbf{A}^{(N)}}) \otimes \cdots \otimes (\widehat{\mathbf{A}}^{(n+1)\top} \widehat{\mathbf{A}}^{(n+1)} + M^{(n+1)} \widehat{\Sigma}_{\mathbf{A}^{(n+1)}}) \right. \right. \right. \\ &\quad \otimes (\widehat{\mathbf{A}}^{(n-1)\top} \widehat{\mathbf{A}}^{(n-1)} + M^{(n-1)} \widehat{\Sigma}_{\mathbf{A}^{(n-1)}}) \otimes \cdots \otimes (\widehat{\mathbf{A}}^{(1)\top} \widehat{\mathbf{A}}^{(1)} + M^{(1)} \widehat{\Sigma}_{\mathbf{A}^{(1)}}) \left. \right) \mathbf{G}_{(n)}^\top \Big\rangle_{r(\mathcal{G})} \\ &\quad \left. + \sigma^2 \mathbf{C}_{\mathbf{A}^{(n)}}^{-1} \right)^{-1}. \end{aligned} \quad (3.79)$$

The expectation in Eqs. (3.79) is explicitly given by

$$\begin{aligned} &\left\langle \mathbf{G}_{(n)} \left((\widehat{\mathbf{A}}^{(N)\top} \widehat{\mathbf{A}}^{(N)} + M^{(N)} \widehat{\Sigma}_{\mathbf{A}^{(N)}}) \otimes \cdots \otimes (\widehat{\mathbf{A}}^{(n+1)\top} \widehat{\mathbf{A}}^{(n+1)} + M^{(n+1)} \widehat{\Sigma}_{\mathbf{A}^{(n+1)}}) \right. \right. \\ &\quad \otimes (\widehat{\mathbf{A}}^{(n-1)\top} \widehat{\mathbf{A}}^{(n-1)} + M^{(n-1)} \widehat{\Sigma}_{\mathbf{A}^{(n-1)}}) \otimes \cdots \otimes (\widehat{\mathbf{A}}^{(1)\top} \widehat{\mathbf{A}}^{(1)} + M^{(1)} \widehat{\Sigma}_{\mathbf{A}^{(1)}}) \left. \right) \mathbf{G}_{(n)}^\top \Big\rangle_{r(\mathcal{G})} \\ &= \widehat{\mathbf{G}}_{(n)} \left((\widehat{\mathbf{A}}^{(N)\top} \widehat{\mathbf{A}}^{(N)} + M^{(N)} \widehat{\Sigma}_{\mathbf{A}^{(N)}}) \otimes \cdots \otimes (\widehat{\mathbf{A}}^{(n+1)\top} \widehat{\mathbf{A}}^{(n+1)} + M^{(n+1)} \widehat{\Sigma}_{\mathbf{A}^{(n+1)}}) \right. \\ &\quad \otimes (\widehat{\mathbf{A}}^{(n-1)\top} \widehat{\mathbf{A}}^{(n-1)} + M^{(n-1)} \widehat{\Sigma}_{\mathbf{A}^{(n-1)}}) \otimes \cdots \otimes (\widehat{\mathbf{A}}^{(1)\top} \widehat{\mathbf{A}}^{(1)} + M^{(1)} \widehat{\Sigma}_{\mathbf{A}^{(1)}}) \left. \right) \widehat{\mathbf{G}}_{(n)}^\top + \boldsymbol{\Xi}^{(n)}, \end{aligned} \quad (3.80)$$

where the entries of $\boldsymbol{\Xi}^{(n)} \in \mathbb{R}^{H^{(n)} \times H^{(n)}}$ are specified as

$$\begin{aligned} \boldsymbol{\Xi}_{h^{(n)}, h'^{(n)}}^{(n)} &= \sum_{(h^{(1)}, h'^{(1)}), \dots, (h^{(n-1)}, h'^{(n-1)}), (h^{(n+1)}, h'^{(n+1)}), \dots, (h^{(N)}, h'^{(N)})} (\widehat{\mathbf{A}}^{(N)\top} \widehat{\mathbf{A}}^{(N)} + M^{(N)} \widehat{\Sigma}_{\mathbf{A}^{(N)}})_{h^{(N)}, h'^{(N)}} \\ &\quad \cdots (\widehat{\mathbf{A}}^{(n+1)\top} \widehat{\mathbf{A}}^{(n+1)} + M^{(n+1)} \widehat{\Sigma}_{\mathbf{A}^{(n+1)}})_{h^{(n+1)}, h'^{(n+1)}} (\widehat{\mathbf{A}}^{(n-1)\top} \widehat{\mathbf{A}}^{(n-1)} + M^{(n-1)} \widehat{\Sigma}_{\mathbf{A}^{(n-1)}})_{h^{(n-1)}, h'^{(n-1)}} \\ &\quad \cdots (\widehat{\mathbf{A}}^{(1)\top} \widehat{\mathbf{A}}^{(1)} + M^{(1)} \widehat{\Sigma}_{\mathbf{A}^{(1)}})_{h^{(1)}, h'^{(1)}} (\boldsymbol{\Sigma}_{\mathcal{G}})_{(h^{(1)}, h'^{(1)}), \dots, (h^{(N)}, h'^{(N)})}. \end{aligned}$$

Here, we used the tensor expression of $\boldsymbol{\Sigma}_{\mathcal{G}} \in \mathbb{R}^{\prod_{n=1}^N 2H^{(n)}}$ for the core posterior covariance $\widehat{\Sigma}_{\mathcal{G}}$.

Free Energy as a Function of Variational Parameters

Substituting Eq. (3.75) into Eq. (3.68), we have

$$\begin{aligned} 2F &= 2 \left(\log r(\mathcal{G}) + \sum_{n=1}^N \log r(\mathbf{A}^{(n)}) \right. \\ &\quad \left. - \log p(\mathcal{V}|\mathcal{G}, \{\mathbf{A}^{(n)}\}) p(\mathcal{G}) \prod_{n=1}^N p(\mathbf{A}^{(n)}) \right)_{r(\mathcal{G}) \left(\prod_{n=1}^N r(\mathbf{A}^{(n)}) \right)} \end{aligned}$$

$$\begin{aligned}
&= \left(\prod_{n=1}^N M^{(n)} \right) \log(2\pi\sigma^2) + \log \frac{\det(\check{\mathbf{C}}_G)}{\det(\widehat{\Sigma}_G)} + \sum_{n=1}^N M^{(n)} \log \frac{\det(C_{A^{(n)}})}{\det(\widehat{\Sigma}_{A^{(n)}})} \\
&\quad + \frac{\|\mathcal{V}\|^2}{\sigma^2} - \prod_{n=1}^N H^{(n)} - \prod_{n=1}^N (M^{(n)} H^{(n)}) \\
&\quad + \text{tr} \left(\check{\mathbf{C}}_G^{-1} (\check{\mathbf{g}} \check{\mathbf{g}}^\top + \widehat{\Sigma}_G) \right) + \sum_{n=1}^N \text{tr} \left(C_{A^{(n)}}^{-1} (\widehat{A}^{(n)\top} \widehat{A}^{(n)} + M^{(n)} \widehat{\Sigma}_{A^{(n)}}) \right) \\
&\quad - \frac{2}{\sigma^2} \check{\mathbf{v}}^\top (\widehat{A}^{(N)} \otimes \cdots \otimes \widehat{A}^{(1)}) \check{\mathbf{g}} \\
&\quad + \frac{1}{\sigma^2} \text{tr} \left\{ \left((\widehat{A}^{(N)\top} \widehat{A}^{(N)} + M^{(N)} \widehat{\Sigma}_{A^{(N)}}) \otimes \cdots \otimes (\widehat{A}^{(1)\top} \widehat{A}^{(1)} + M^{(1)} \widehat{\Sigma}_{A^{(1)}}) \right) \right. \\
&\quad \left. \cdot (\check{\mathbf{g}} \check{\mathbf{g}}^\top + \widehat{\Sigma}_G) \right\}. \tag{3.81}
\end{aligned}$$

Empirical Variational Bayesian Algorithm

The derivative of the free energy (3.81) with respect to $\check{\mathbf{C}}_G$ gives

$$2 \frac{\partial F}{\partial \check{\mathbf{C}}_G} = M^{(n)} \left(\check{\mathbf{C}}_G^{-1} - \check{\mathbf{C}}_G^{-2} \left(\check{\mathbf{g}} \check{\mathbf{g}}^\top + \widehat{\Sigma}_G \right) \right).$$

Since it holds that

$$\frac{\partial \check{\mathbf{C}}_G}{\partial (C_{G^{(n)}})_{h,h}} = \mathbf{C}_{G^{(N)}} \otimes \cdots \otimes \mathbf{C}_{G^{(n+1)}} \otimes E_{(H^{(n)}, h, h)} \otimes \mathbf{C}_{G^{(n-1)}} \otimes \cdots \otimes \mathbf{C}_{G^{(1)}},$$

where $E_{(H, h, h')} \in \mathbb{R}^{H \times H}$ is the matrix with the (h, h') th entry equal to one and the others equal to zero, we have

$$\begin{aligned}
2 \frac{\partial F}{\partial (C_{G^{(n)}})_{h,h}} &= 2 \text{tr} \left(\frac{\partial F}{\partial \check{\mathbf{C}}_G} \frac{\partial \check{\mathbf{C}}_G}{\partial (C_{G^{(n)}})_{h,h}} \right) \\
&= M^{(n)} \left\| \text{vec} \left(I_{H^{(N)}} \otimes \cdots \otimes I_{H^{(n+1)}} \otimes (C_{G^{(n)}})_{h,h}^{-1} E_{(H^{(n)}, h, h)} \otimes I_{H^{(n-1)}} \otimes \cdots \otimes I_{H^{(1)}} \right. \right. \\
&\quad \left. \left. - C_{G^{(N)}}^{-1} \otimes \cdots \otimes C_{G^{(n+1)}}^{-1} \otimes (C_{G^{(n)}})_{h,h}^{-2} E_{(H^{(n)}, h, h)} \otimes C_{G^{(n-1)}}^{-1} \otimes \cdots \otimes C_{G^{(1)}}^{-1} (\check{\mathbf{g}} \check{\mathbf{g}}^\top + \widehat{\Sigma}_G) \right) \right\|_1 \\
&= M^{(n)} (C_{G^{(n)}})_{h,h}^{-2} \left\| \text{vec} \left(I_{H^{(N)}} \otimes \cdots \otimes I_{H^{(n+1)}} \otimes (C_{G^{(n)}})_{h,h} E_{(H^{(n)}, h, h)} \otimes I_{H^{(n-1)}} \otimes \cdots \otimes I_{H^{(1)}} \right. \right. \\
&\quad \left. \left. - C_{G^{(N)}}^{-1} \otimes \cdots \otimes C_{G^{(n+1)}}^{-1} \otimes E_{(H^{(n)}, h, h)} \otimes C_{G^{(n-1)}}^{-1} \otimes \cdots \otimes C_{G^{(1)}}^{-1} (\check{\mathbf{g}} \check{\mathbf{g}}^\top + \widehat{\Sigma}_G) \right) \right\|_1 \\
&= M^{(n)} (C_{G^{(n)}})_{h,h}^{-2} \left(\left(\prod_{n' \neq n} H^{(n')} \right) (C_{G^{(n)}})_{h,h} - \text{diag} \left(\check{\mathbf{g}} \check{\mathbf{g}}^\top + \widehat{\Sigma}_G \right)^\top \right. \\
&\quad \left. \text{diag} \left(C_{G^{(N)}}^{-1} \otimes \cdots \otimes C_{G^{(n+1)}}^{-1} \otimes E_{(H^{(n)}, h, h)} \otimes C_{G^{(n-1)}}^{-1} \otimes \cdots \otimes C_{G^{(1)}}^{-1} \right) \right), \tag{3.82}
\end{aligned}$$

Algorithm 3 EVB learning for Tucker factorization.

-
- 1: Initialize the variational parameters $(\widehat{\mathbf{g}}, \widehat{\Sigma}_G, \{\widehat{\mathbf{A}}^{(n)}\}, \{\widehat{\Sigma}_{A^{(n)}}\})$, and the hyper-parameters $(\{\mathbf{C}_{G^{(n)}}\}, \{\mathbf{C}_{A^{(n)}}\}, \sigma^2)$, for example, $\widehat{\mathbf{g}}_h \sim \text{Gauss}_1(0, \tau)$, $\widehat{\mathbf{A}}_{m,h}^{(n)} \sim \text{Gauss}_1(0, \tau^{1/N})$, $\widehat{\Sigma}_G = \tau \mathbf{I}_{\prod_{n=1}^N H^{(n)}}$, $\mathbf{C}_{G^{(n)}} = \tau \mathbf{I}_{H^{(n)}}$, $\widehat{\Sigma}_{A^{(n)}} = \mathbf{C}_{A^{(n)}} = \tau^{1/N} \mathbf{I}_{H^{(n)}}$, and $\sigma^2 = \tau^2$ for $\tau^2 = \|\mathcal{V}\|^2 / \prod_{n=1}^N M^{(n)}$.
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (3.77), (3.76), (3.79), and (3.78) to update $\widehat{\Sigma}_G$, $\widehat{\mathbf{g}}$, $\{\widehat{\Sigma}_{A^{(n)}}\}$, and $\{\widehat{\mathbf{A}}^{(n)}\}$, respectively.
 - 3: Apply Eqs. (3.83) through (3.85) to update $\{\mathbf{C}_{G^{(n)}}\}$, $\{\mathbf{C}_{A^{(n)}}\}$, and σ^2 , respectively.
 - 4: Prune the h th component if $c_{g_h^{(n)}}^2 c_{a_h^{(n)}}^2 < \varepsilon$, where $\varepsilon > 0$ is a threshold, e.g., set to $\varepsilon = 10^{-4}$.
 - 5: Evaluate the free energy (3.81).
 - 6: Iterate Steps 2 through 5 until convergence (until the energy decrease becomes smaller than a threshold).
-

where $\|\cdot\|_1$ denotes the ℓ_1 -norm, and $\text{diag}(\cdot)$ denotes the column vector consisting of the diagonal entries of a matrix. Thus, the prior covariance for the core tensor can be updated by

$$c_{g_h^{(n)}}^2 = \frac{\text{diag}(\widehat{\mathbf{g}} \widehat{\mathbf{g}}^\top + \widehat{\Sigma}_G)^\top \text{diag}(C_{G^{(N)}}^{-1} \otimes \cdots \otimes C_{G^{(n+1)}}^{-1} \otimes E_{(H^{(n)}, h, h)} \otimes C_{G^{(n-1)}}^{-1} \otimes \cdots \otimes C_{G^{(1)}}^{-1})}{\prod_{n' \neq n} H^{(n')}}. \quad (3.83)$$

The derivative of the free energy (3.81) with respect to $\mathbf{C}_{A^{(n)}}$ gives

$$2 \frac{\partial F}{\partial c_{a_h^{(n)}}^2} = M^{(n)} \left(c_{a_h^{(n)}}^{-2} - c_{a_h^{(n)}}^{-4} \left(\frac{\|\widehat{\mathbf{a}}_h^{(n)}\|^2}{M^{(n)}} + (\widehat{\Sigma}_{A^{(n)}})_{h,h} \right) \right),$$

which leads to the following update rule for the prior covariance for the factor matrices:

$$c_{a_h^{(n)}}^2 = \frac{\|\widehat{\mathbf{a}}_h^{(n)}\|^2}{M^{(n)}} + (\widehat{\Sigma}_{A^{(n)}})_{h,h}. \quad (3.84)$$

Finally, the derivative of the free energy (3.81) with respect to σ^2 gives

$$\begin{aligned} 2 \frac{\partial F}{\partial \sigma^2} &= \frac{\prod_{n=1}^N M^{(n)}}{\sigma^2} - \frac{1}{\sigma^4} \left(\|\mathcal{V}\|^2 - 2 \check{\mathbf{v}}^\top (\widehat{\mathbf{A}}^{(N)} \otimes \cdots \otimes \widehat{\mathbf{A}}^{(1)}) \widehat{\mathbf{g}} \right. \\ &\quad \left. + \text{tr} \left\{ \left((\widehat{\mathbf{A}}^{(N)\top} \widehat{\mathbf{A}}^{(N)} + M^{(N)} \widehat{\Sigma}_{A^{(N)}}) \otimes \cdots \otimes (\widehat{\mathbf{A}}^{(1)\top} \widehat{\mathbf{A}}^{(1)} + M^{(1)} \widehat{\Sigma}_{A^{(1)}}) \right) (\widehat{\check{\mathbf{g}}} \widehat{\check{\mathbf{g}}}^\top + \widehat{\Sigma}_G) \right\} \right), \end{aligned}$$

which leads to the update rule for the noise variance as follows:

$$\begin{aligned}\widehat{\sigma}^2 = & \frac{1}{\prod_{n=1}^N M^{(n)}} \left(\|\mathcal{V}\|^2 - 2\tilde{\mathbf{v}}^\top (\widehat{\mathbf{A}}^{(N)} \otimes \cdots \otimes \widehat{\mathbf{A}}^{(1)}) \widehat{\mathbf{g}} \right. \\ & \left. + \text{tr} \left\{ \left((\widehat{\mathbf{A}}^{(N)\top} \widehat{\mathbf{A}}^{(N)} + M^{(N)} \widehat{\Sigma}_{\mathbf{A}^{(N)}}) \otimes \cdots \otimes (\widehat{\mathbf{A}}^{(1)\top} \widehat{\mathbf{A}}^{(1)} + M^{(1)} \widehat{\Sigma}_{\mathbf{A}^{(1)}}) \right) (\widehat{\mathbf{g}} \widehat{\mathbf{g}}^\top + \widehat{\Sigma}_G) \right\} \right).\end{aligned}\quad (3.85)$$

Algorithm 3 summarizes the EVB algorithm for Tucker factorization. If we appropriately set the hyperparameters ($\{\mathbf{C}_{G^{(n)}}\}, \{\mathbf{C}_{A^{(n)}}\}, \sigma^2$) in Step 1 and skip Steps 3 and 4, Algorithm 3 is reduced to (nonempirical) VB learning.

3.4 Low-Rank Subspace Clustering

PCA *globally* embeds data points into a low-dimensional subspace. As more flexible tools, *subspace clustering* methods, which *locally* embed the data into the union of subspaces, have been developed. In this section, we derive VB learning for subspace clustering.

3.4.1 Subspace Clustering Methods

Most clustering methods, such as k -means (MacQueen, 1967; Lloyd, 1982) and spectral clustering (Shi and Malik, 2000), (explicitly or implicitly) assume that there are sparse areas between dense areas, and separate the dense areas as clusters (Figure 3.2 left). On the other hand, there are some data, e.g., projected trajectories of points on a rigid body in 3D space, where data points can be assumed to lie in a union of small dimensional subspaces (Figure 3.2 right). Note that a point lying in a subspace is not necessarily far from a point lying in another subspace if those subspaces intersect each other. Subspace clustering methods have been developed to analyze this kind of data.



Figure 3.2 Clustering (left) and subspace clustering (right).

Let $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_M) \in \mathbb{R}^{L \times M}$ be L -dimensional observed samples of size M . We assume that each \mathbf{v}_m is approximately expressed as a linear combination of M' words in a dictionary, $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_{M'}) \in \mathbb{R}^{L \times M'}$, i.e.,

$$\mathbf{V} = \mathbf{DU} + \mathbf{\mathcal{E}},$$

where $\mathbf{U} \in \mathbb{R}^{M' \times M}$ is unknown coefficients, and $\mathbf{\mathcal{E}} \in \mathbb{R}^{L \times M}$ is noise. In subspace clustering, the observed matrix \mathbf{V} itself is often used as a dictionary \mathbf{D} . Then, a convex formulation of the *sparse subspace clustering (SSC)* (Soltanolkotabi and Candès, 2011; Elhamifar and Vidal, 2013) is given by

$$\min_{\mathbf{U}} \|\mathbf{V} - \mathbf{V}\mathbf{U}\|_{\text{Fro}}^2 + \lambda \|\mathbf{U}\|_1, \text{ s.t. } \text{diag}(\mathbf{U}) = \mathbf{0}, \quad (3.86)$$

where $\mathbf{U} \in \mathbb{R}^{M \times M}$ is a parameter to be estimated, $\lambda > 0$ is a regularization coefficient. $\|\cdot\|_1$ is the ℓ_1 -norm of a matrix. The first term in Eq. (3.86) together with the constraint requires that each data point \mathbf{v}_m is accurately expressed as a linear combination of other data points, $\{\mathbf{v}_{m'}\}$ for $m' \neq m$. The second term, which is the ℓ_1 -regularizer, enforces that the number of samples contributing to the linear combination should be small, which leads to low-dimensionality of each obtained subspace. After the solution $\widehat{\mathbf{U}}$ to the problem (3.86) is obtained, the matrix $\text{abs}(\widehat{\mathbf{U}}) + \text{abs}(\widehat{\mathbf{U}}^\top)$, where $\text{abs}(\cdot)$ takes the absolute value elementwise, is regarded as an affinity matrix, and a *spectral clustering algorithm*, such as the *normalized cuts* (Shi and Malik, 2000), is applied to obtain clusters.

Another popular method for subspace clustering is *low-rank subspace clustering (LRSC)* or *low-rank representation* (Liu et al., 2010; Liu and Yan, 2011; Liu et al., 2012; Vidal and Favaro, 2014), where low-dimensional subspaces are sought by enforcing the low-rankness of \mathbf{U} with the *trace norm*:

$$\min_{\mathbf{U}} \|\mathbf{V} - \mathbf{V}\mathbf{U}\|_{\text{Fro}}^2 + \lambda \|\mathbf{U}\|_{\text{tr}}. \quad (3.87)$$

Since LRSC enforces the low-rankness of \mathbf{U} , the constraint $\text{diag}(\mathbf{U}) = \mathbf{0}$ is not necessary, which makes its optimization problem (3.87) significantly simpler than the optimization problem (3.86) for SSC. Thanks to this simplicity, the global solution of Eq. (3.87) has been analytically obtained (Vidal and Favaro, 2014).

Good properties of SSC and LRSC have been theoretically shown (Liu et al., 2010, 2012; Soltanolkotabi and Candès, 2011; Elhamifar and Vidal, 2013; Vidal and Favaro, 2014). It is observed that they behave differently in different situations, and each of SSC and LRSC shows advantages and disadvantages over the other, i.e., neither SSC nor LRSC is known to dominate the other in the general situations. In the rest of this section, we focus on LRSC and derive its VB learning algorithm.

3.4.2 VB Learning for LRSC

We start with the following probabilistic model, of which the maximum a posteriori (MAP) estimator coincides with the solution to the convex formulation (3.87) under a certain hyperparameter setting:

$$p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{V} - \mathbf{V}\mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2\right), \quad (3.88)$$

$$p(\mathbf{A}) \propto \exp\left(-\frac{1}{2} \text{tr}(\mathbf{A}\mathbf{C}_A^{-1}\mathbf{A}^\top)\right), \quad (3.89)$$

$$p(\mathbf{B}) \propto \exp\left(-\frac{1}{2} \text{tr}(\mathbf{B}\mathbf{C}_B^{-1}\mathbf{B}^\top)\right). \quad (3.90)$$

Here, we factorized \mathbf{U} as $\mathbf{U} = \mathbf{B}\mathbf{A}^\top$, where $\mathbf{A} \in \mathbb{R}^{M \times H}$ and $\mathbf{B} \in \mathbb{R}^{M \times H}$ for $H \leq \min(L, M)$ are the parameters to be estimated (Babacan et al., 2012a). This factorization is known to induce low-rankness through the MIR mechanism, which will be discussed in Chapter 7. We assume that the prior covariances are diagonal:

$$\mathbf{C}_A = \mathbf{Diag}(c_{a_1}^2, \dots, c_{a_H}^2), \quad \mathbf{C}_B = \mathbf{Diag}(c_{b_1}^2, \dots, c_{b_H}^2).$$

Conditional Conjugacy

The model likelihood (3.88) of LRSC is similar to the model likelihood (3.1) of MF, and it is in the Gaussian form with respect to \mathbf{A} if \mathbf{B} is regarded as a constant, or vice versa. Therefore, the priors (3.89) and (3.90) are conditionally conjugate for \mathbf{A} and \mathbf{B} , respectively.

Variational Bayesian Algorithm

The conditional conjugacy implies that the following independence constraint on the approximate posterior leads to a tractable algorithm:

$$r(\mathbf{A}, \mathbf{B}) = r(\mathbf{A})r(\mathbf{B}).$$

In the same way as MF, we can show that the VB posterior is Gaussian in the following form:

$$\begin{aligned} r(\mathbf{A}) &\propto \exp\left(-\frac{\text{tr}\left((\mathbf{A} - \widehat{\mathbf{A}})\widehat{\Sigma}_A^{-1}(\mathbf{A} - \widehat{\mathbf{A}})^\top\right)}{2}\right), \\ r(\mathbf{B}) &\propto \exp\left(-\frac{(\check{\mathbf{b}} - \widehat{\mathbf{b}})^\top \widehat{\Sigma}_B^{-1}(\check{\mathbf{b}} - \widehat{\mathbf{b}})}{2}\right), \end{aligned} \quad (3.91)$$

where $\check{\mathbf{b}} = \text{vec}(\mathbf{B}) \in \mathbb{R}^{MH}$. The variational parameters satisfy the following stationary conditions:

$$\widehat{\mathbf{A}} = \frac{1}{\sigma^2} \mathbf{V}^\top \mathbf{V} \widehat{\mathbf{B}} \widehat{\Sigma}_A, \quad (3.92)$$

$$\widehat{\Sigma}_A = \sigma^2 \left(\langle \mathbf{B}^\top \mathbf{V}^\top \mathbf{V} \mathbf{B} \rangle_{r(\mathbf{B})} + \sigma^2 \mathbf{C}_A^{-1} \right)^{-1}, \quad (3.93)$$

$$\widehat{\mathbf{b}} = \frac{\widehat{\Sigma}_B}{\sigma^2} \text{vec}(\mathbf{V}^\top \mathbf{V} \widehat{\mathbf{A}}), \quad (3.94)$$

$$\widehat{\Sigma}_B = \sigma^2 \left((\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A) \otimes \mathbf{V}^\top \mathbf{V} + \sigma^2 (\mathbf{C}_B^{-1} \otimes \mathbf{I}_M) \right)^{-1}, \quad (3.95)$$

where the entries of $\langle \mathbf{B}^\top \mathbf{V}^\top \mathbf{V} \mathbf{B} \rangle_{r(\mathbf{B})}$ in Eq. (3.93) are explicitly given by

$$\left(\langle \mathbf{B}^\top \mathbf{V}^\top \mathbf{V} \mathbf{B} \rangle_{r(\mathbf{B})} \right)_{h,h'} = \left(\mathbf{B}^\top \mathbf{V}^\top \mathbf{V} \widehat{\mathbf{B}} \right)_{h,h'} + \text{tr} \left(\mathbf{V}^\top \mathbf{V} \widehat{\Sigma}_B^{(h,h')} \right). \quad (3.96)$$

Here $\widehat{\Sigma}_B^{(h,h')} \in \mathbb{R}^{M \times M}$ is the (h, h') th block matrix of $\widehat{\Sigma}_B \in \mathbb{R}^{MH \times MH}$, i.e.,

$$\widehat{\Sigma}_B = \begin{pmatrix} \widehat{\Sigma}_B^{(1,1)} & \cdots & \widehat{\Sigma}_B^{(1,H)} \\ \vdots & \ddots & \vdots \\ \widehat{\Sigma}_B^{(H,1)} & \cdots & \widehat{\Sigma}_B^{(H,H)} \end{pmatrix}.$$

Free Energy as a Function of Variational Parameters

The free energy can be explicitly written as

$$\begin{aligned} 2F = LM \log(2\pi\sigma^2) + \frac{\| \mathbf{V} - \mathbf{V} \widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top \|_{\text{Fro}}^2}{\sigma^2} + M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + \log \frac{\det(\mathbf{C}_B \otimes \mathbf{I}_M)}{\det(\widehat{\Sigma}_B)} \\ - 2MH + \text{tr} \left\{ \mathbf{C}_A^{-1} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) \right\} + \text{tr} \left\{ \mathbf{C}_B^{-1} \widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} \right\} + \text{tr} \left\{ (\mathbf{C}_B^{-1} \otimes \mathbf{I}_M) \widehat{\Sigma}_B \right\} \\ + \text{tr} \left\{ \sigma^{-2} \mathbf{V}^\top \mathbf{V} \left(-\widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\mathbf{B}}^\top + \left\langle \mathbf{B} (\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A) \mathbf{B}^\top \right\rangle_{r(\mathbf{B})} \right) \right\}, \end{aligned} \quad (3.97)$$

where the expectation in the last term is given by

$$\begin{aligned} \left(\left\langle \mathbf{B} (\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A) \mathbf{B} \right\rangle_{r(\mathbf{B})}^\top \right)_{m,m'} = & \left(\widehat{\mathbf{B}} (\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A) \widehat{\mathbf{B}}^\top \right)_{m,m'} \\ & + \text{tr} \left((\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A) \widehat{\Sigma}_B^{(m,m')} \right). \end{aligned} \quad (3.98)$$

Algorithm 4 EVB learning for low-rank subspace clustering.

-
- 1: Initialize the variational parameters $(\widehat{\mathbf{A}}, \widehat{\Sigma}_A, \widehat{\mathbf{B}}, \widehat{\Sigma}_B)$, and the hyperparameters $(\mathbf{C}_A, \mathbf{C}_B, \sigma^2)$, for example, $\widehat{\mathbf{A}}_{m,h}, \widehat{\mathbf{B}}_{m,h} \sim \text{Gauss}_1(0, 1^2)$, $\widehat{\Sigma}_A = \mathbf{C}_A = \mathbf{C}_B = \mathbf{I}_H$, $\widehat{\Sigma}_B = \mathbf{I}_{MH}$, and $\sigma^2 = \|V\|^2/(LM)$.
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (3.93), (3.92), (3.95), and (3.94) to update $\widehat{\Sigma}_A$, $\widehat{\mathbf{A}}$, $\widehat{\Sigma}_B$, and $\widehat{\mathbf{B}}$, respectively.
 - 3: Apply Eqs. (3.99) through (3.101) to update \mathbf{C}_A , \mathbf{C}_B , and σ^2 , respectively.
 - 4: Prune the h th component if $c_{ah}^2 c_{bh}^2 < \varepsilon$, where $\varepsilon > 0$ is a threshold, e.g., set to $\varepsilon = 10^{-4}$.
 - 5: Evaluate the free energy (3.97).
 - 6: Iterate Steps 2 through 5 until convergence (until the energy decrease becomes smaller than a threshold).
-

Here, $\widehat{\Sigma}_B^{(m,m')} \in \mathbb{R}^{H \times H}$ is defined as

$$\widehat{\Sigma}_B^{(m,m')} = \begin{pmatrix} \left(\widehat{\Sigma}_B^{(1,1)}\right)_{m,m'} & \cdots & \left(\widehat{\Sigma}_B^{(1,H)}\right)_{m,m'} \\ \vdots & \ddots & \vdots \\ \left(\widehat{\Sigma}_B^{(H,1)}\right)_{m,m'} & \cdots & \left(\widehat{\Sigma}_B^{(H,H)}\right)_{m,m'} \end{pmatrix}.$$

Empirical Variational Bayesian Algorithm

By differentiating the free energy (3.97) with respect to the hyperparameters, we can obtain the stationary conditions for the hyperparameters:

$$c_{ah}^2 = \|\widehat{\mathbf{a}}_h\|^2/M + (\widehat{\Sigma}_A)_{h,h}, \quad (3.99)$$

$$c_{bh}^2 = \left(\|\widehat{\mathbf{b}}_h\|^2 + \text{tr}(\widehat{\Sigma}_B^{(h,h)})\right)/M, \quad (3.100)$$

$$\widehat{\sigma}^2 = \frac{\text{tr}\left(V^\top V \left(\mathbf{I}_M - 2\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top + \left\langle \mathbf{B}(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M\widehat{\Sigma}_A)\mathbf{B}^\top\right\rangle_{r(\mathbf{B})}\right)\right)}{LM}. \quad (3.101)$$

Algorithm 4 summarizes the EVB algorithm. If we appropriately set the hyperparameters $(\mathbf{C}_A, \mathbf{C}_B, \sigma^2)$ in Step 1 and skip Steps 3 and 4, Algorithm 4 is reduced to (nonempirical) VB learning.

Variational Bayesian Algorithm under the Kronecker Product Covariance Constraint

The standard VB algorithm, given in Algorithm 4, for LRSC requires the inversion of an $MH \times MH$ matrix, which is prohibitively huge in practical

applications. As a remedy, Babacan et al. (2012a) proposed to restrict the posterior $r(\mathbf{B})$ for \mathbf{B} to be the matrix variate Gaussian with the Kronecker product covariance (KPC) structure, as Eq. (3.18). Namely, we restrict the approximate posterior to be in the following form:

$$\begin{aligned} r(\mathbf{B}) &= \text{MGauss}_{M,H}(\mathbf{B}; \widehat{\mathbf{B}}, \widehat{\boldsymbol{\Psi}}_B \otimes \widehat{\boldsymbol{\Sigma}}_B) \\ &\propto \exp\left(-\frac{1}{2}\text{tr}\left(\widehat{\boldsymbol{\Psi}}_B^{-1}(\mathbf{B} - \widehat{\mathbf{B}})\widehat{\boldsymbol{\Sigma}}_B^{-1}(\mathbf{B} - \widehat{\mathbf{B}})^\top\right)\right). \end{aligned} \quad (3.102)$$

Under this additional constraint, the free energy is written as

$$\begin{aligned} 2F^{\text{KPC}} &= LM \log(2\pi\sigma^2) + \frac{\|V - V\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top\|^2}{\sigma^2} + M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\boldsymbol{\Sigma}}_A)} + M \log \frac{\det(\mathbf{C}_B)}{\det(\widehat{\boldsymbol{\Sigma}}_B)} \\ &\quad + H \log \frac{1}{\det(\widehat{\boldsymbol{\Psi}}_B)} - 2MH \\ &\quad + \text{tr}\left\{\mathbf{C}_A^{-1}\left(\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\boldsymbol{\Sigma}}_A\right)\right\} + \text{tr}\left\{\mathbf{C}_B^{-1}\left(\widehat{\mathbf{B}}^\top\widehat{\mathbf{B}} + \text{tr}(\widehat{\boldsymbol{\Psi}}_B)\widehat{\boldsymbol{\Sigma}}_B\right)\right\} \\ &\quad + \text{tr}\left\{\sigma^{-2}V^\top V\left(M\widehat{\mathbf{B}}\widehat{\boldsymbol{\Sigma}}_A\widehat{\mathbf{B}}^\top + \text{tr}\left(\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\boldsymbol{\Sigma}}_A\right)\widehat{\boldsymbol{\Sigma}}_B\right)\widehat{\boldsymbol{\Psi}}_B\right\}. \end{aligned} \quad (3.103)$$

By differentiating the free energy (3.103) with respect to each variational parameter, we obtain the following update rules:

$$\widehat{\mathbf{A}} = \frac{1}{\sigma^2}V^\top V\widehat{\mathbf{B}}\widehat{\boldsymbol{\Sigma}}_A, \quad (3.104)$$

$$\widehat{\boldsymbol{\Sigma}}_A = \sigma^2\left(\widehat{\mathbf{B}}^\top V^\top V\widehat{\mathbf{B}} + \text{tr}(V^\top V\widehat{\boldsymbol{\Psi}}_B)\widehat{\boldsymbol{\Sigma}}_B + \sigma^2\mathbf{C}_A^{-1}\right)^{-1}, \quad (3.105)$$

$$\widehat{\mathbf{B}}^{\text{new}} = \widehat{\mathbf{B}}^{\text{old}} - \alpha\left(\widehat{\mathbf{B}}^{\text{old}}\mathbf{C}_B^{-1} + \frac{1}{\sigma^2}V^\top V\left(-\widehat{\mathbf{A}} + \widehat{\mathbf{B}}^{\text{old}}(\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\boldsymbol{\Sigma}}_A)\right)\right), \quad (3.106)$$

$$\widehat{\boldsymbol{\Sigma}}_B = \sigma^2\left(\frac{\text{tr}(V^\top V\widehat{\boldsymbol{\Psi}}_B)}{M}(\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\boldsymbol{\Sigma}}_A) + \frac{\sigma^2\text{tr}(\widehat{\boldsymbol{\Psi}}_B)}{M}\mathbf{C}_B^{-1}\right)^{-1}, \quad (3.107)$$

$$\widehat{\boldsymbol{\Psi}}_B = \sigma^2\left(\frac{\text{tr}((\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\boldsymbol{\Sigma}}_A)\widehat{\boldsymbol{\Sigma}}_B)}{H}V^\top V + \frac{\sigma^2\text{tr}(\mathbf{C}_B^{-1}\widehat{\boldsymbol{\Sigma}}_B)}{H}\mathbf{I}_M\right)^{-1}, \quad (3.108)$$

$$c_{a_h}^2 = (\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\boldsymbol{\Sigma}}_A)_{h,h}/M, \quad (3.109)$$

$$c_{b_h}^2 = \left(\widehat{\mathbf{B}}^\top\widehat{\mathbf{B}} + \text{tr}(\widehat{\boldsymbol{\Psi}}_B)\widehat{\boldsymbol{\Sigma}}_B\right)_{h,h}/M, \quad (3.110)$$

$$\begin{aligned} \sigma^2 &= \frac{1}{LM}\text{tr}\left(V^\top V\left(\mathbf{I}_M - 2\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top + \text{tr}((\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\boldsymbol{\Sigma}}_A)\widehat{\boldsymbol{\Sigma}}_B)\widehat{\boldsymbol{\Psi}}_B\right.\right. \\ &\quad \left.\left.+\widehat{\mathbf{B}}(\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\boldsymbol{\Sigma}}_A)\widehat{\mathbf{B}}^\top\right)\right). \end{aligned} \quad (3.111)$$

Algorithm 5 EVB learning for low-rank subspace clustering under the Kronecker product covariance constraint (3.102).

- 1: Initialize the variational parameters $(\widehat{\mathbf{A}}, \widehat{\Sigma}_A, \widehat{\mathbf{B}}, \widehat{\Sigma}_B, \widehat{\Psi}_B)$, and the hyperparameters $(\mathbf{C}_A, \mathbf{C}_B, \sigma^2)$, for example, $\widehat{\mathbf{A}}_{m,h}, \widehat{\mathbf{B}}_{m,h} \sim \text{Gauss}_1(0, 1^2)$, $\widehat{\Sigma}_A = \widehat{\Sigma}_B = \mathbf{C}_A = \mathbf{C}_B = \mathbf{I}_H$, $\widehat{\Psi}_B = \mathbf{I}_M$, and $\sigma^2 = \|\mathbf{V}\|^2/(LM)$.
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (3.105), (3.104), (3.107), and (3.108) to update $\widehat{\Sigma}_A$, $\widehat{\mathbf{A}}$, $\widehat{\Sigma}_B$, and $\widehat{\Psi}_B$, respectively.
 - 3: Apply Eq. (3.106) T times (e.g., $T = 20$) to update $\widehat{\mathbf{B}}$.
 - 4: Apply Eqs. (3.109) through (3.111) to update \mathbf{C}_A , \mathbf{C}_B , and σ^2 , respectively.
 - 5: Prune the h th component if $c_{a_h}^2 c_{b_h}^2 < \varepsilon$, where $\varepsilon > 0$ is a threshold, e.g., set to $\varepsilon = 10^{-4}$.
 - 6: Evaluate the free energy (3.103).
 - 7: Iterate Steps 2 through 6 until convergence (until the energy decrease becomes smaller than a threshold).
-

Note that Eq. (3.106) is the gradient descent algorithm for $\widehat{\mathbf{B}}$ with the step size $\alpha > 0$.

Algorithm 5 summarizes the EVB algorithm under the KPC constraint, which we call *KPC approximation (KPCA)*. If we appropriately set the hyperparameters $(\mathbf{C}_A, \mathbf{C}_B, \sigma^2)$ in Step 1 and skip Steps 4 and 5, Algorithm 5 is reduced to (nonempirical) VB learning.

3.5 Sparse Additive Matrix Factorization

PCA is known to be sensitive to outliers in data and generally fails in their presence. To cope with outliers, robust PCA, where spiky noise is captured by an elementwise sparse term, was proposed (Candès et al., 2011). In this section, we introduce a generalization of robust PCA, called *sparse additive matrix factorization (SAMF)* (Nakajima et al., 2013b) and derive its VB learning algorithm.

3.5.1 Robust PCA and Matrix Factorization

In *robust PCA*, the observed matrix $\mathbf{V} \in \mathbb{R}^{L \times M}$ is modeled as follows:

$$\mathbf{V} = \mathbf{U}^{\text{low-rank}} + \mathbf{U}^{\text{element}} + \mathcal{E}, \quad (3.112)$$

where $\mathbf{U}^{\text{low-rank}} \in \mathbb{R}^{L \times M}$ is a low-rank matrix, $\mathbf{U}^{\text{element}} \in \mathbb{R}^{L \times M}$ is an elementwise sparse matrix, and $\mathcal{E} \in \mathbb{R}^{L \times M}$ is a (typically dense) noise matrix.

Given the observed matrix \mathbf{V} , one can infer each term in the right-hand side of Eq. (3.112) by solving the following convex problem (Candès et al., 2011):

$$\min_{\mathbf{U}^{\text{low-rank}}, \mathbf{U}^{\text{element}}} \|\mathbf{V} - \mathbf{U}^{\text{low-rank}} - \mathbf{U}^{\text{element}}\|_{\text{Fro}}^2 + \lambda_1 \|\mathbf{U}^{\text{low-rank}}\|_{\text{tr}} + \lambda_2 \|\mathbf{U}^{\text{element}}\|_1,$$

where the *trace norm* $\|\cdot\|_{\text{tr}}$ induces low-rank sparsity, and the ℓ_1 -norm $\|\cdot\|_1$ induces elementwise sparsity. The regularization coefficients λ_1 and λ_2 control the strength of sparsity.

In Bayesian modeling, the low-rank matrix is commonly expressed as the product of two matrices, $\mathbf{A} \in \mathbb{R}^{M \times H}$ and $\mathbf{B} \in \mathbb{R}^{L \times H}$:

$$\mathbf{U}^{\text{low-rank}} = \mathbf{B}\mathbf{A}^\top = \sum_{h=1}^H \mathbf{b}_h \mathbf{a}_h^\top. \quad (3.113)$$

Trivially, low-rankness is forced if H is set to a small value. However, when VB learning is applied, the estimator can be low-rank even if we adopt the full-rank model, i.e., $H = \min(L, M)$. This phenomenon is caused by MIR, which will be discussed in Chapter 7.

3.5.2 Sparse Matrix Factorization Terms

SAMF (Nakajima et al., 2013b) was proposed as a generalization of robust PCA, where various types of sparsity are induced by combining different types of factorization. For example, the following factorization *implicitly* induces rowwise, columnwise, and elementwise sparsity, respectively:

$$\mathbf{U}^{\text{row}} = \boldsymbol{\Gamma}_E \mathbf{D} = (\gamma_1^e \tilde{\mathbf{d}}_1, \dots, \gamma_L^e \tilde{\mathbf{d}}_L)^\top, \quad (3.114)$$

$$\mathbf{U}^{\text{column}} = \mathbf{E} \boldsymbol{\Gamma}_D = (\gamma_1^d \mathbf{e}_1, \dots, \gamma_M^d \mathbf{e}_M), \quad (3.115)$$

$$\mathbf{U}^{\text{element}} = \mathbf{E} \odot \mathbf{D}, \quad (3.116)$$

where $\boldsymbol{\Gamma}_D = \text{Diag}(\gamma_1^d, \dots, \gamma_M^d) \in \mathbb{R}^{M \times M}$ and $\boldsymbol{\Gamma}_E = \text{Diag}(\gamma_1^e, \dots, \gamma_L^e) \in \mathbb{R}^{L \times L}$ are diagonal matrices, and $\mathbf{D}, \mathbf{E} \in \mathbb{R}^{L \times M}$. \odot denotes the Hadamard product, i.e., $(\mathbf{E} \odot \mathbf{D})_{l,m} = E_{l,m} D_{l,m}$. The reason why the factorizations (3.114) through (3.116) induce the corresponding types of sparsity is explained in Section 7.5.

As a general expression of sparsity inducing factorizations, we define a sparse matrix factorization (SMF) term with a mapping \mathbf{G} consisting of partitioning, rearrangement, and factorization:

$$\mathbf{U} = \mathbf{G}(\{\mathbf{U}'^{(k)}\}_{k=1}^K; \mathcal{X}), \text{ where } \mathbf{U}'^{(k)} = \mathbf{B}^{(k)} \mathbf{A}^{(k)\top}. \quad (3.117)$$

Here, $\{\mathbf{A}^{(k)}, \mathbf{B}^{(k)}\}_{k=1}^K$ are parameters to be estimated, and $\mathbf{G}(\cdot; \mathcal{X}) : \mathbb{R}^{\prod_{k=1}^K (L'^{(k)} \times M'^{(k)})} \mapsto \mathbb{R}^{L \times M}$ is a designed function associated with an index mapping \mathcal{X} (explained shortly).

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_{1,1} & \mathbf{U}_{1,2} & \mathbf{U}_{1,3} & \mathbf{U}_{1,4} \\ \mathbf{U}_{2,1} & \mathbf{U}_{2,2} & \mathbf{U}_{2,3} & \mathbf{U}_{2,4} \\ \mathbf{U}_{3,1} & \mathbf{U}_{3,2} & \mathbf{U}_{3,3} & \mathbf{U}_{3,4} \\ \mathbf{U}_{4,1} & \mathbf{U}_{4,2} & \mathbf{U}_{4,3} & \mathbf{U}_{4,4} \end{pmatrix} \xleftarrow{\mathbf{G}} \begin{array}{l} \mathbf{U}'^{(1)} = \begin{pmatrix} \mathbf{U}_{1,1} & \mathbf{U}_{1,2} & \mathbf{U}_{1,3} & \mathbf{U}_{1,4} \end{pmatrix} = \mathbf{B}^{(1)} \mathbf{A}^{(1)\top} \\ \mathbf{U}'^{(2)} = \begin{pmatrix} \mathbf{U}_{2,1} & \mathbf{U}_{2,2} \\ \mathbf{U}_{3,1} & \mathbf{U}_{3,2} \end{pmatrix} = \mathbf{B}^{(2)} \mathbf{A}^{(2)\top} \\ \mathbf{U}'^{(3)} = \begin{pmatrix} \mathbf{U}_{2,3} & \mathbf{U}_{2,4} & \mathbf{U}_{3,3} & \mathbf{U}_{3,4} \end{pmatrix} = \mathbf{B}^{(3)} \mathbf{A}^{(3)\top} \\ \mathbf{U}'^{(4)} = \begin{pmatrix} \mathbf{U}_{4,1} & \mathbf{U}_{4,2} & \mathbf{U}_{4,3} \end{pmatrix} = \mathbf{B}^{(4)} \mathbf{A}^{(4)\top} \\ \mathbf{U}'^{(5)} = \begin{pmatrix} \mathbf{U}_{4,4} \end{pmatrix} = \mathbf{B}^{(5)} \mathbf{A}^{(5)\top} \end{array}$$

Figure 3.3 An example of SMF-term construction. $\mathbf{G}(\cdot; \mathcal{X})$ with $\mathcal{X} : (k, l', m') \mapsto (l, m)$ maps the set $\{\mathbf{U}'^{(k)}\}_{k=1}^K$ of the PR matrices to the target matrix \mathbf{U} , so that $\mathbf{U}'_{l',m'} = U_{\mathcal{X}(k,l',m')} = U_{l,m}$.

Figure 3.3 illustrates how to construct an SMF term. First, we partition the entries of \mathbf{U} into K parts. Then, by rearranging the entries in each part, we form *partitioned-and-rearranged (PR) matrices* $\mathbf{U}'^{(k)} \in \mathbb{R}^{L'^{(k)} \times M'^{(k)}}$ for $k = 1, \dots, K$. Finally, each of $\mathbf{U}'^{(k)}$ is decomposed into the product of $\mathbf{A}^{(k)} \in \mathbb{R}^{M'^{(k)} \times H'^{(k)}}$ and $\mathbf{B}^{(k)} \in \mathbb{R}^{L'^{(k)} \times H'^{(k)}}$, where $H'^{(k)} \leq \min(L'^{(k)}, M'^{(k)})$.

In Eq. (3.117), the function $\mathbf{G}(\cdot; \mathcal{X})$ is responsible for partitioning and rearrangement: it maps the set $\{\mathbf{U}'^{(k)}\}_{k=1}^K$ of the PR matrices to the target matrix $\mathbf{U} \in \mathbb{R}^{L \times M}$, based on the one-to-one map $\mathcal{X} : (k, l', m') \mapsto (l, m)$ from the indices of the entries in $\{\mathbf{U}'^{(k)}\}_{k=1}^K$ to the indices of the entries in \mathbf{U} such that

$$\left(\mathbf{G}(\{\mathbf{U}'^{(k)}\}_{k=1}^K; \mathcal{X}) \right)_{l,m} = U_{l,m} = U_{\mathcal{X}(k,l',m')} = U'_{l',m'}. \quad (3.118)$$

When VB learning is applied, the SMF-term expression (3.117) induces partitionwise sparsity and low-rank sparsity in each partition. Accordingly, partitioning, rearrangement, and factorization should be designed in the following way. Suppose that we are given a required sparsity structure on a matrix (examples of possible side information that suggests particular sparsity structures are given in Section 3.5.3). We first partition the matrix, according to the required sparsity. Some partitions can be submatrices. We rearrange each of the submatrices on which we do not want to impose low-rank sparsity into a long vector ($\mathbf{U}'^{(3)}$ in the example in Figure 3.3). We leave the other submatrices which we want to be low-rank ($\mathbf{U}'^{(2)}$) and the original vectors ($\mathbf{U}'^{(1)}$ and $\mathbf{U}'^{(4)}$) and scalars ($\mathbf{U}'^{(5)}$) as they are. Finally, we factorize each of the PR matrices to induce sparsity.

Let us, for example, assume that rowwise sparsity is required. We first make the rowwise partition, i.e., separate $\mathbf{U} \in \mathbb{R}^{L \times M}$ into L pieces of M -dimensional row vectors $\mathbf{U}'^{(l)} = \tilde{\mathbf{u}}_l^\top \in \mathbb{R}^{1 \times M}$. Then, we factorize each partition as $\mathbf{U}'^{(l)} = \mathbf{B}^{(l)} \mathbf{A}^{(l)\top}$ (see the top illustration in Figure 3.4). Thus, we obtain the rowwise sparse term (3.114). Here, $\mathcal{X}(k, 1, m') = (k, m')$ makes the following connection between Eqs. (3.114) and (3.117): $\gamma_l^e = \mathbf{B}^{(k)} \in \mathbb{R}, \bar{\mathbf{d}}_l = \mathbf{A}^{(k)} \in \mathbb{R}^{M \times 1}$ for $k = l$. Similarly, requiring columnwise and elementwise sparsity leads to

Table 3.1 Examples of SMF terms.

Factorization	Induced sparsity	K	$(L'^{(k)}, M'^{(k)})$	$\mathcal{X} : (k, l', m') \mapsto (l, m)$
$\mathbf{U} = \mathbf{B}\mathbf{A}^\top$	low-rank	1	(L, M)	$\mathcal{X}(1, l', m') = (l', m')$
$\mathbf{U} = \mathbf{\Gamma}_E \mathbf{D}$	rowwise	L	$(1, M)$	$\mathcal{X}(k, 1, m') = (k, m')$
$\mathbf{U} = \mathbf{E}\mathbf{\Gamma}_D$	columnwise	M	$(L, 1)$	$\mathcal{X}(k, l', 1) = (l', k)$
$\mathbf{U} = \mathbf{E} \odot \mathbf{D}$	elementwise	$L \times M$	$(1, 1)$	$\mathcal{X}(k, 1, 1) = \text{vec-order}(k)$

$$\begin{aligned}
& \mathbf{U} = \begin{pmatrix} \mathbf{U}_{1,1} & \mathbf{U}_{1,2} & \mathbf{U}_{1,3} \\ \mathbf{U}_{2,1} & \mathbf{U}_{2,2} & \mathbf{U}_{2,3} \end{pmatrix} \xleftarrow{G} \begin{array}{l} \mathbf{U}'^{(1)} = \begin{pmatrix} \mathbf{U}_{1,1} & \mathbf{U}_{1,2} & \mathbf{U}_{1,3} \end{pmatrix} = \mathbf{B}^{(1)} \mathbf{A}^{(1)\top} \\ \mathbf{U}'^{(2)} = \begin{pmatrix} \mathbf{U}_{2,1} & \mathbf{U}_{2,2} & \mathbf{U}_{2,3} \end{pmatrix} = \mathbf{B}^{(2)} \mathbf{A}^{(2)\top} \end{array} \\
& \mathbf{U} = \begin{pmatrix} \mathbf{U}_{1,1} \\ \mathbf{U}_{2,1} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{1,2} \\ \mathbf{U}_{2,2} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{1,3} \\ \mathbf{U}_{2,3} \end{pmatrix} \xleftarrow{G} \begin{array}{l} \mathbf{U}'^{(1)} = \begin{pmatrix} \mathbf{U}_{1,1} \\ \mathbf{U}_{2,1} \end{pmatrix} = \mathbf{B}^{(1)} \mathbf{A}^{(1)\top} \\ \mathbf{U}'^{(2)} = \begin{pmatrix} \mathbf{U}_{1,2} \\ \mathbf{U}_{2,2} \end{pmatrix} = \mathbf{B}^{(2)} \mathbf{A}^{(2)\top} \end{array} \quad \mathbf{U}'^{(3)} = \begin{pmatrix} \mathbf{U}_{1,3} \\ \mathbf{U}_{2,3} \end{pmatrix} = \mathbf{B}^{(3)} \mathbf{A}^{(3)\top} \\
& \mathbf{U} = \begin{pmatrix} \mathbf{U}_{1,1} \\ \mathbf{U}_{2,1} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{1,2} \\ \mathbf{U}_{2,2} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{1,3} \\ \mathbf{U}_{2,3} \end{pmatrix} \xleftarrow{G} \begin{array}{l} \mathbf{U}'^{(1)} = \begin{pmatrix} \mathbf{U}_{1,1} \\ \mathbf{U}_{2,1} \end{pmatrix} = \mathbf{B}^{(1)} \mathbf{A}^{(1)\top} \quad \mathbf{U}'^{(4)} = \begin{pmatrix} \mathbf{U}_{2,2} \\ \mathbf{U}_{2,3} \end{pmatrix} = \mathbf{B}^{(4)} \mathbf{A}^{(4)\top} \\ \mathbf{U}'^{(2)} = \begin{pmatrix} \mathbf{U}_{1,2} \\ \mathbf{U}_{2,2} \end{pmatrix} = \mathbf{B}^{(2)} \mathbf{A}^{(2)\top} \quad \mathbf{U}'^{(5)} = \begin{pmatrix} \mathbf{U}_{1,3} \\ \mathbf{U}_{2,3} \end{pmatrix} = \mathbf{B}^{(5)} \mathbf{A}^{(5)\top} \\ \mathbf{U}'^{(3)} = \begin{pmatrix} \mathbf{U}_{1,1} \\ \mathbf{U}_{1,2} \end{pmatrix} = \mathbf{B}^{(3)} \mathbf{A}^{(3)\top} \quad \mathbf{U}'^{(6)} = \begin{pmatrix} \mathbf{U}_{1,3} \\ \mathbf{U}_{2,3} \end{pmatrix} = \mathbf{B}^{(6)} \mathbf{A}^{(6)\top} \end{array}
\end{aligned}$$

Figure 3.4 SMF-term construction for the rowwise (top), the columnwise (middle), and the elementwise (bottom) sparse terms.

Eqs. (3.115) and (3.116), respectively (see the bottom two illustrations in Figure 3.4). Table 3.1 summarizes how to design these SMF terms, where $\text{vec-order}(k) = (1 + ((k - 1) \bmod L), \lceil k/L \rceil)$ goes along the columns one after another in the same way as the vec operator forming a vector by stacking the columns of a matrix (in other words, $(\mathbf{U}'^{(1)}, \dots, \mathbf{U}'^{(K)})^\top = \text{vec}(\mathbf{U})$).

Now we define the SAMF model as the sum of SMF terms (3.117):

$$\mathbf{V} = \sum_{s=1}^S \mathbf{U}^{(s)} + \mathcal{E}, \quad (3.119)$$

$$\text{where } \mathbf{U}^{(s)} = \mathbf{G}(\{\mathbf{B}^{(k,s)} \mathbf{A}^{(k,s)\top}\}_{k=1}^{K^{(s)}}; \mathcal{X}^{(s)}). \quad (3.120)$$

3.5.3 Examples of SMF Terms

In practice, SMF terms should be designed based on side information. Suppose that $\mathbf{V} \in \mathbb{R}^{L \times M}$ consists of M samples of L -dimensional sensor outputs. In robust PCA (3.112), we add an elementwise sparse term (3.116) to the low-rank term (3.113), assuming that the low-rank signal is expected to be



Figure 3.5 Foreground/background video separation task.

contaminated with spiky noise when observed. Here, we can say that the existence of spiky noise is used as side information.

Similarly, if we expect that a small number of sensors can be broken, and their outputs are unreliable over all M samples, we should add the rowwise sparse term (3.114) to separate the low-rank signal from rowwise noise:

$$\mathbf{V} = \mathbf{U}^{\text{low-rank}} + \mathbf{U}^{\text{row}} + \mathcal{E}.$$

If we expect some accidental disturbances occurred during the observation, but do not know their exact locations (i.e., which samples are affected), the columnwise sparse term (3.115) can effectively capture such disturbances.

The SMF expression (3.117) enables us to use side information in a more flexible way, and its advantage has been shown in a *foreground/background video separation* problem (Nakajima et al., 2013b). The top image in Figure 3.5 is a frame of a video available from the *Caviar Project* website,¹ and the task is to separate *moving* objects (bottom-right) from the background (bottom-left). Previous approaches (Candès et al., 2011; Ding et al., 2011; Babacan et al., 2012b) first constructed the observation matrix \mathbf{V} by stacking all pixels in each frame into each column (Figure 3.6), and then fitted it by the robust PCA model (3.112). Here, the low-rank term and the elementwise

¹ The European Commission (EC)-funded CAVIAR project/IST 2001 37540, found at URL: <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>.

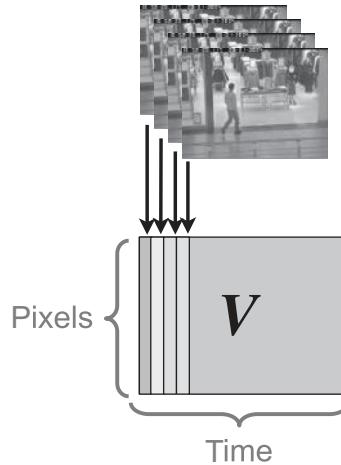


Figure 3.6 The observation matrix V is constructed by stacking all pixels in each frame into each column.

sparse term are expected to capture the static background and the moving foreground, respectively. However, we can also rely on the natural assumption that the pixels in a segment sharing similar intensities tend to belong to the same object. Under this assumption as side information, we can adopt a segmentwise sparse term, for which the PR matrix is constructed based on a precomputed oversegmented image (Figure 3.7). The segmentwise sparse term has been shown to capture the foreground more accurately than the elementwise sparse term in this application. Details will be discussed in Chapter 11.

3.5.4 VB Learning for SAMF

Let us summarize the parameters of the SAMF model (3.119) as follows:

$$\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_A^{(s)}, \boldsymbol{\Theta}_B^{(s)}\}_{s=1}^S, \quad \text{where} \quad \boldsymbol{\Theta}_A^{(s)} = \{A^{(k,s)}\}_{k=1}^{K^{(s)}}, \quad \boldsymbol{\Theta}_B^{(s)} = \{B^{(k,s)}\}_{k=1}^{K^{(s)}}.$$

As in the MF model, we assume independent Gaussian noise and priors. Then, the likelihood and the priors are given by

$$p(V|\boldsymbol{\Theta}) \propto \exp\left(-\frac{1}{2\sigma^2} \left\| V - \sum_{s=1}^S U^{(s)} \right\|_{\text{Fro}}^2\right), \quad (3.121)$$

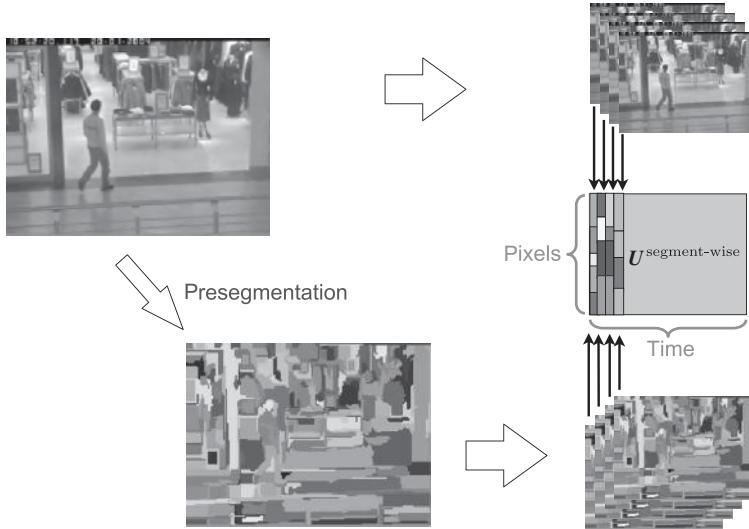


Figure 3.7 Construction of a segmentwise sparse term. The original frame is presegmented, based on which the segmentwise sparse term is constructed as an SMF term.

$$p(\{\boldsymbol{\Theta}_A^{(s)}\}_{s=1}^S) \propto \exp \left(-\frac{1}{2} \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr} \left(\mathbf{A}^{(k,s)} \mathbf{C}_A^{(k,s)-1} \mathbf{A}^{(k,s)\top} \right) \right), \quad (3.122)$$

$$p(\{\boldsymbol{\Theta}_B^{(s)}\}_{s=1}^S) \propto \exp \left(-\frac{1}{2} \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr} \left(\mathbf{B}^{(k,s)} \mathbf{C}_B^{(k,s)-1} \mathbf{B}^{(k,s)\top} \right) \right). \quad (3.123)$$

We assume that the prior covariances of $\mathbf{A}^{(k,s)}$ and $\mathbf{B}^{(k,s)}$ are diagonal:

$$\mathbf{C}_A^{(k,s)} = \text{Diag}(c_{a_1}^{(k,s)2}, \dots, c_{a_H}^{(k,s)2}),$$

$$\mathbf{C}_B^{(k,s)} = \text{Diag}(c_{b_1}^{(k,s)2}, \dots, c_{b_H}^{(k,s)2}).$$

Conditional Conjugacy

As seen in Eq. (3.121), the SAMF model is the MF model for the parameters $(\boldsymbol{\Theta}_A^{(s)}, \boldsymbol{\Theta}_B^{(s)})$ in the s th SMF term, if the other parameters $\{\boldsymbol{\Theta}_A^{(s')}, \boldsymbol{\Theta}_B^{(s')}\}_{s' \neq s}$ are regarded as constants. Therefore, the Gaussian priors (3.122) and (3.123) are conditionally conjugate for each of $\boldsymbol{\Theta}_A^{(s)}$ and $\boldsymbol{\Theta}_B^{(s)}$ in each of the SMF terms.

Variational Bayesian Algorithm

Based on the conditional conjugacy, we solve the VB learning problem under the following independence constraint (Babacan et al., 2012b):

$$r(\boldsymbol{\Theta}) = \prod_{s=1}^S r_A^{(s)}(\boldsymbol{\Theta}_A^{(s)}) r_B^{(s)}(\boldsymbol{\Theta}_B^{(s)}). \quad (3.124)$$

Following the standard procedure described in Section 2.1.5, we can find that the VB posterior, which minimizes the free energy (2.15), is in the following form:

$$\begin{aligned} r(\boldsymbol{\Theta}) &= \prod_{s=1}^S \prod_{k=1}^{K^{(s)}} \left(\text{MGauss}_{M'^{(k,s)}, H'^{(k,s)}}(\mathbf{A}^{(k,s)}; \widehat{\mathbf{A}}^{(k,s)}, \widehat{\Sigma}_A^{(k,s)}) \right. \\ &\quad \cdot \left. \text{MGauss}_{L'^{(k,s)}, H'^{(k,s)}}(\mathbf{B}^{(k,s)}; \widehat{\mathbf{B}}^{(k,s)}, \widehat{\Sigma}_B^{(k,s)}) \right) \\ &= \prod_{s=1}^S \prod_{k=1}^{K^{(s)}} \left(\prod_{m'=1}^{M'^{(k,s)}} \text{Gauss}_{H'^{(k,s)}}(\widehat{\mathbf{a}}_{m'}^{(k,s)}; \widetilde{\widehat{\mathbf{a}}}_{m'}^{(k,s)}, \widehat{\Sigma}_A^{(k,s)}) \right. \\ &\quad \cdot \left. \prod_{l'=1}^{L'^{(k,s)}} \text{Gauss}_{H'^{(k,s)}}(\widehat{\mathbf{b}}_{l'}^{(k,s)}; \widetilde{\widehat{\mathbf{b}}}_{l'}^{(k,s)}, \widehat{\Sigma}_B^{(k,s)}) \right) \end{aligned} \quad (3.125)$$

with the variational parameters satisfying the stationary conditions given by

$$\widehat{\mathbf{A}}^{(k,s)} = \sigma^{-2} \mathbf{Z}'^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} \widehat{\Sigma}_A^{(k,s)}, \quad (3.126)$$

$$\widehat{\Sigma}_A^{(k,s)} = \sigma^2 \left(\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\Sigma}_B^{(k,s)} + \sigma^2 \mathbf{C}_A^{(k,s)-1} \right)^{-1}, \quad (3.127)$$

$$\widehat{\mathbf{B}}^{(k,s)} = \sigma^{-2} \mathbf{Z}'^{(k,s)} \widehat{\mathbf{A}}^{(k,s)} \widehat{\Sigma}_B^{(k,s)}, \quad (3.128)$$

$$\widehat{\Sigma}_B^{(k,s)} = \sigma^2 \left(\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\Sigma}_A^{(k,s)} + \sigma^2 \mathbf{C}_B^{(k,s)-1} \right)^{-1}. \quad (3.129)$$

Here, $\mathbf{Z}'^{(k,s)} \in \mathbb{R}^{L'^{(k,s)} \times M'^{(k,s)}}$ is defined as

$$\mathbf{Z}'^{(k,s)} = \mathbf{Z}_{\chi^{(s)}(k, l', m')}^{(s)}, \quad \text{where } \mathbf{Z}^{(s)} = \mathbf{V} - \sum_{s' \neq s} \widehat{\mathbf{U}}^{(s)}. \quad (3.130)$$

Free Energy as a Function of Variational Parameters

The free energy can be explicitly written as

$$\begin{aligned} 2F &= LM \log(2\pi\sigma^2) + \frac{\|\mathbf{V}\|_{\text{Fro}}^2}{\sigma^2} \\ &\quad + \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \left(M'^{(k,s)} \log \frac{\det(\mathbf{C}_A^{(k,s)})}{\det(\widehat{\Sigma}_A^{(k,s)})} + L'^{(k,s)} \log \frac{\det(\mathbf{C}_B^{(k,s)})}{\det(\widehat{\Sigma}_B^{(k,s)})} \right) \\ &\quad + \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr} \left\{ \mathbf{C}_A^{(k,s)-1} (\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\Sigma}_A^{(k,s)}) \right. \\ &\quad \left. + \mathbf{C}_B^{(k,s)-1} (\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\Sigma}_B^{(k,s)}) \right\} \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{\sigma^2} \text{tr} \left\{ -2V^\top \left(\sum_{s=1}^S \mathbf{G}(\{\widehat{\mathbf{B}}^{(k,s)} \widehat{\mathbf{A}}^{(k,s)\top}\}_{k=1}^{K^{(s)}}; \mathcal{X}^{(s)}) \right) \right. \\
& \quad \left. + 2 \sum_{s=1}^S \sum_{s'=s+1}^S \mathbf{G}^\top(\{\widehat{\mathbf{B}}^{(k,s)} \widehat{\mathbf{A}}^{(k,s)\top}\}_{k=1}^{K^{(s)}}; \mathcal{X}^{(s')}) \mathbf{G}(\{\widehat{\mathbf{B}}^{(k,s')} \widehat{\mathbf{A}}^{(k,s')\top}\}_{k=1}^{K^{(s')}}; \mathcal{X}^{(s')}) \right\} \\
& + \frac{1}{\sigma^2} \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr} \left\{ (\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\Sigma}_A^{(k,s)}) (\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\Sigma}_B^{(k,s)}) \right\} \\
& - \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} (L'^{(k,s)} + M'^{(k,s)}) H'^{(k,s)}. \tag{3.131}
\end{aligned}$$

Empirical Variational Bayesian Algorithm

The following stationary conditions for the hyperparameters can be obtained from the derivatives of the free energy (3.131):

$$c_{a_h}^{(k,s)2} = \left\| \widehat{\mathbf{a}}_h^{(k,s)} \right\|^2 / M'^{(k,s)} + (\widehat{\Sigma}_A^{(k,s)})_{hh}, \tag{3.132}$$

$$c_{b_h}^{(k,s)2} = \left\| \widehat{\mathbf{b}}_h^{(k,s)} \right\|^2 / L'^{(k,s)} + (\widehat{\Sigma}_B^{(k,s)})_{hh}, \tag{3.133}$$

Algorithm 6 EVB learning for sparse additive matrix factorization.

- 1: Initialize the variational parameters $\{\widehat{\mathbf{A}}^{(k,s)}, \widehat{\Sigma}_A^{(k,s)}, \widehat{\mathbf{B}}^{(k,s)}, \widehat{\Sigma}_B^{(k,s)}\}_{k=1,s=1}^{K^{(s)} S}$, and the hyperparameters $\{\mathbf{C}_A^{(k,s)}, \mathbf{C}_B^{(k,s)}\}_{k=1,s=1}^{K^{(s)} S}, \sigma^2$, for example, $\widehat{\mathbf{A}}_{m,h}^{(k,s)}, \widehat{\mathbf{B}}_{l,h}^{(k,s)} \sim \text{Gauss}_1(0, \tau), \widehat{\Sigma}_A^{(k,s)} = \widehat{\Sigma}_B^{(k,s)} = \mathbf{C}_A^{(k,s)} = \mathbf{C}_B^{(k,s)} = \tau \mathbf{I}_{H'^{(k,s)}},$ and $\sigma^2 = \tau^2$ for $\tau^2 = \|V\|_{\text{Fro}}^2 / (LM)$.
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (3.127), (3.126), (3.129), and (3.128) for each k and s to update $\widehat{\Sigma}_A^{(k,s)}, \widehat{\mathbf{A}}^{(k,s)}, \widehat{\Sigma}_B^{(k,s)}$, and $\widehat{\mathbf{B}}^{(k,s)}$, respectively.
 - 3: Apply Eqs. (3.132) and (3.133) for all $h' = 1, \dots, H'^{(k,s)}$, k and s , and Eq. (3.134) to update $\mathbf{C}_A^{(k,s)}, \mathbf{C}_B^{(k,s)}$, and σ^2 , respectively.
 - 4: Prune the h th component if $c_{a_h}^{(k,s)2} c_{b_h}^{(k,s)2} < \varepsilon$, where $\varepsilon > 0$ is a threshold, e.g., set to $\varepsilon = 10^{-4}$.
 - 5: Evaluate the free energy (3.131).
 - 6: Iterate Steps 2 through 5 until convergence (until the energy decrease becomes smaller than a threshold).
-

$$\begin{aligned} \sigma^2 = & \frac{1}{LM} \left\{ \|V\|_{\text{Fro}}^2 - 2 \sum_{s=1}^S \text{tr} \left(\widehat{\mathbf{U}}^{(s)\top} \left(V - \sum_{s'=s+1}^S \widehat{\mathbf{U}}^{(s')} \right) \right) \right. \\ & \left. + \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr} \left((\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\boldsymbol{\Sigma}}_B^{(k,s)}) \cdot (\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\boldsymbol{\Sigma}}_A^{(k,s)}) \right) \right\}. \end{aligned} \quad (3.134)$$

Algorithm 6 summarizes the EVB algorithm for SAMF. If we appropriately set the hyperparameters $\{\mathbf{C}_A^{(k,s)}, \mathbf{C}_B^{(k,s)}\}_{k=1,s=1}^{K^{(s)} S}, \sigma^2$ in Step 1 and skip Steps 3 and 4, Algorithm 6 is reduced to (nonempirical) VB learning.

4

VB Algorithm for Latent Variable Models

In this chapter, we discuss VB learning for *latent variable models*. Starting with finite mixture models as the simplest example, we overview the VB learning algorithms for more complex latent variable models such as Bayesian networks and hidden Markov models.

Let \mathcal{H} denote the set of (*local*) *latent variables* and \mathbf{w} denote a model parameter vector (or the set of *global latent variables*). In this chapter, we consider the latent variable model for training data \mathcal{D} :

$$p(\mathcal{D}|\mathbf{w}) = \sum_{\mathcal{H}} p(\mathcal{D}, \mathcal{H}|\mathbf{w}).$$

Let us employ the following factorized model to approximate the posterior distribution for \mathbf{w} and \mathcal{H} :

$$r(\mathbf{w}, \mathcal{H}) = r_w(\mathbf{w})r_{\mathcal{H}}(\mathcal{H}). \quad (4.1)$$

Applying the general VB framework explained in Section 2.1.5 to the preceding model leads to the following update rules for \mathbf{w} and \mathcal{H} :

$$r_w(\mathbf{w}) = \frac{1}{C_w} p(\mathbf{w}) \exp \langle \log p(\mathcal{D}, \mathcal{H}|\mathbf{w}) \rangle_{r_{\mathcal{H}}(\mathcal{H})}, \quad (4.2)$$

$$r_{\mathcal{H}}(\mathcal{H}) = \frac{1}{C_{\mathcal{H}}} \exp \langle \log p(\mathcal{D}, \mathcal{H}|\mathbf{w}) \rangle_{r_w(\mathbf{w})}. \quad (4.3)$$

In the following sections, we discuss these update rules for some specific examples of latent variable models.

4.1 Finite Mixture Models

A *finite mixture model* $p(\mathbf{x}|\mathbf{w})$ of an L -dimensional input $\mathbf{x} \in \mathbb{R}^L$ with a parameter vector $\mathbf{w} \in \mathbb{R}^M$ is defined by

$$p(\mathbf{x}|\mathbf{w}) = \sum_{k=1}^K \alpha_k p(\mathbf{x}|\boldsymbol{\tau}_k), \quad (4.4)$$

where integer K is the number of components and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\top \in \Delta^{K-1}$ is the set of mixing weights (Example 1.3). The parameter \mathbf{w} of the model is $\mathbf{w} = \{\alpha_k, \boldsymbol{\tau}_k\}_{k=1}^K$.

The finite mixture model can be rewritten as follows by using a hidden variable $\mathbf{z} = (z_1, \dots, z_K)^\top \in \{\mathbf{e}_1, \dots, \mathbf{e}_K\}$,

$$p(\mathbf{x}, \mathbf{z}|\mathbf{w}) = \prod_{k=1}^K [\alpha_k p(\mathbf{x}|\boldsymbol{\tau}_k)]^{z_k}. \quad (4.5)$$

Here $\mathbf{e}_k \in \{0, 1\}^K$ is the K -dimensional binary vector, called the *one-of- K representation*, with one at the k th entry and zeros at the other entries:

$$\mathbf{e}_k = (\underbrace{0, \dots, 0, \underbrace{1}_{\text{k-th}}, 0, \dots, 0}_{K}, 0, \dots, 0)^\top.$$

The hidden variable \mathbf{z} is not observed and is representing the component from which the data sample \mathbf{x} is generated. If the data sample \mathbf{x} is from the k th component, then $z_k = 1$, otherwise, $z_k = 0$. Then

$$\sum_z p(\mathbf{x}, \mathbf{z}|\mathbf{w}) = p(\mathbf{x}|\mathbf{w})$$

holds where the sum over \mathbf{z} goes through all possible values of the hidden variable.

4.1.1 Mixture of Gaussians

If the component distribution in Eq. (4.4) is chosen to be a Gaussian distribution,

$$p(\mathbf{x}|\boldsymbol{\tau}) = \text{Gauss}_L(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

the finite mixture model is called the *mixture of Gaussians* or the *Gaussian mixture model (GMM)*.

In some applications, the parameters are restricted to the means of each component, and it is assumed that there is no correlation between each input dimension. In this case, since $L = M$, the model is written by

$$p(\mathbf{x}|\mathbf{w}) = \sum_{k=1}^K \frac{\alpha_k}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_k\|^2}{2\sigma^2}\right), \quad (4.6)$$

where $\sigma > 0$ is a constant.

In this subsection, the uncorrelated GMM (4.6) is considered in the VB framework by further assuming that $\sigma^2 = 1$ for simplicity. The joint model for the observed and hidden variables (4.5) is given by the product of the following two distributions:

$$p(\mathbf{z}|\boldsymbol{\alpha}) = \text{Multinomial}_{K,1}(\mathbf{z}; \boldsymbol{\alpha}), \quad (4.7)$$

$$p(\mathbf{x}|\mathbf{z}, \{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{k=1}^K \{\text{Gauss}_M(\mathbf{x}; \boldsymbol{\mu}_k, \mathbf{I}_M)\}^{z_k}. \quad (4.8)$$

Thus, for the set of hidden variables $\mathcal{H} = \{\mathbf{z}^{(n)}\}_{n=1}^N$ and the complete data set $\{\mathcal{D}, \mathcal{H}\} = \{\mathbf{x}^{(n)}, \mathbf{z}^{(n)}\}_{n=1}^N$, the complete likelihood is given by

$$p(\mathcal{D}, \mathcal{H}|\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{n=1}^N \prod_{k=1}^K \left\{ \alpha_k \text{Gauss}_M(\mathbf{x}^{(n)}; \boldsymbol{\mu}_k, \mathbf{I}_M) \right\}^{z_k^{(n)}}. \quad (4.9)$$

ML learning of the GMM is carried out by the expectation-maximization (EM) algorithm (Dempster et al., 1977), which corresponds to a clustering algorithm called the soft K-means (MacKay, 2003, ch. 22).

Because of the conditional conjugacy (Section 2.1.2) of the parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\top \in \Delta^{K-1}$ and $\{\boldsymbol{\mu}_k\}_{k=1}^K$, we assume that the prior of the parameters is the product of the following two distributions:

$$p(\boldsymbol{\alpha}|\phi) = \text{Dirichlet}_K(\boldsymbol{\alpha}; (\phi, \dots, \phi)^\top), \quad (4.10)$$

$$p(\{\boldsymbol{\mu}_k\}_{k=1}^K | \boldsymbol{\mu}_0, \xi) = \prod_{k=1}^K \text{Gauss}_M(\boldsymbol{\mu}_k | \boldsymbol{\mu}_0, (1/\xi)\mathbf{I}_M), \quad (4.11)$$

where $\xi > 0$, $\boldsymbol{\mu}_0 \in \mathbb{R}^M$ and $\phi > 0$ are the hyperparameters.

VB Posterior for the Gaussian Mixture Model

Let

$$\bar{N}_k = \sum_{n=1}^N \langle z_k^{(n)} \rangle_{r_{\mathcal{H}}(\mathcal{H})} \quad (4.12)$$

and

$$\bar{\mathbf{x}}_k = \frac{1}{\bar{N}_k} \sum_{n=1}^N \langle z_k^{(n)} \rangle_{r_{\mathcal{H}}(\mathcal{H})} \mathbf{x}^{(n)}, \quad (4.13)$$

where $z_k^{(n)} = 1$ if the n th data sample $\mathbf{x}^{(n)}$ is from the k th component; otherwise, $z_k^{(n)} = 0$. The variable \bar{N}_k is the expected number of times data come from the k th component, and $\bar{\mathbf{x}}_k$ is the mean of them. Note that the variables \bar{N}_k and $\bar{\mathbf{x}}_k$ satisfy the constraints $\sum_{k=1}^K \bar{N}_k = N$ and $\sum_{k=1}^K \bar{N}_k \bar{\mathbf{x}}_k = \sum_{n=1}^N \mathbf{x}^{(n)}$. From

(4.2) and the respective priors (4.10) and (4.11), the VB posterior $r_w(\boldsymbol{w}) = r_\alpha(\boldsymbol{\alpha})r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K)$ is obtained as the product of the following two distributions:

$$r_\alpha(\boldsymbol{\alpha}) = \text{Dirichlet}_K \left(\boldsymbol{\alpha}; (\widehat{\phi}_1, \dots, \widehat{\phi}_K)^\top \right), \quad (4.14)$$

$$r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{k=1}^K \text{Gauss}_M \left(\boldsymbol{\mu}_k; \widehat{\boldsymbol{\mu}}_k, \widehat{\sigma}_k^2 \mathbf{I}_M \right), \quad (4.15)$$

where

$$\widehat{\phi}_k = \bar{N}_k + \phi, \quad (4.16)$$

$$\widehat{\sigma}_k^2 = \frac{1}{\bar{N}_k + \xi}, \quad (4.17)$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{\bar{N}_k \bar{\mathbf{x}}_k + \xi \boldsymbol{\mu}_0}{\bar{N}_k + \xi}. \quad (4.18)$$

From Eq. (4.3), the VB posterior $r_{\mathcal{H}}(\mathcal{H})$ is given by

$$r_{\mathcal{H}}(\mathcal{H}) = \frac{1}{C_{\mathcal{H}}} \prod_{n=1}^N \exp \left(z_k^{(n)} \left\{ \Psi(\widehat{\phi}_k) - \Psi \left(\sum_{k'=1}^K \widehat{\phi}_{k'} \right) - \frac{\|\mathbf{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\|^2}{2} - \frac{M}{2} \left(\log 2\pi + \frac{1}{\bar{N}_k + \xi} \right) \right\} \right),$$

where Ψ is the di-gamma (psi) function, and we used

$$\langle \log \alpha_k \rangle_{r_\alpha(\boldsymbol{\alpha})} = \Psi(\widehat{\phi}_k) - \Psi \left(\sum_{k'=1}^K \widehat{\phi}_{k'} \right). \quad (4.19)$$

That is, $r_{\mathcal{H}}(\mathcal{H}) = r_z(\{\mathbf{z}^{(n)}\}_{n=1}^N)$ is the multinomial distribution:

$$\begin{aligned} r_z(\{\mathbf{z}^{(n)}\}_{n=1}^N) &= \prod_{n=1}^N r_z(\mathbf{z}^{(n)}) \\ &= \prod_{n=1}^N \text{Multinomial}_{K,1} \left(\mathbf{z}^{(n)}; \bar{\mathbf{z}}^{(n)} \right), \end{aligned}$$

where $\bar{\mathbf{z}}^{(n)} \in \Delta^{K-1}$ is

$$\bar{z}_k^{(n)} = \langle z_k^{(n)} \rangle_{r_{\mathcal{H}}(\mathcal{H})} = \frac{\bar{z}_k^{(n)}}{\sum_{k'=1}^K \bar{z}_{k'}^{(n)}}, \quad (4.20)$$

for

$$\bar{z}_k^{(n)} = \exp \left(\Psi(\widehat{\phi}_k) - \Psi \left(\sum_{k'=1}^K \widehat{\phi}_{k'} \right) - \frac{1}{2} \|\mathbf{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\|^2 + M \widehat{\sigma}_k^2 \right). \quad (4.21)$$

The free energy as a function of variational parameters is expressed as follows:

$$\begin{aligned}
F &= \left\langle \log \frac{r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w})}{p(\mathbf{w})} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w})} - \langle \log p(\mathcal{D}, \mathcal{H}|\mathbf{w}) \rangle_{r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w})} \\
&= \left\langle \log \frac{(\bar{z}_k^{(n)})^{z_k^{(n)}} \frac{\Gamma(\sum_{k=1}^K \widehat{\phi}_k)}{\prod_{k=1}^K \Gamma(\widehat{\phi}_k)} \prod_{k=1}^K \alpha_k^{\widehat{\phi}_k-1} \frac{\exp\left(-\frac{\|\mu_k - \widehat{\mu}_k\|^2}{2\widehat{\sigma}_k^2}\right)}{(2\pi\widehat{\sigma}_k^2)^{M/2}}}{\frac{\Gamma(K\phi)}{(\Gamma(\phi))^K} \prod_{k=1}^K \alpha_k^{\phi-1} \left(\frac{\xi}{2\pi}\right)^{M/2} \exp\left(-\frac{\xi\|\mu_k - \mu_0\|^2}{2}\right)} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w})} \\
&\quad - \left\langle \log \prod_{n=1}^N \prod_{k=1}^K \left\{ \alpha_k \frac{\exp\left(-\frac{\|\mathbf{x}^{(n)} - \mu_k\|^2}{2}\right)}{(2\pi)^{M/2}} \right\}^{z_k^{(n)}} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w})} \\
&= \log \left(\frac{\Gamma(\sum_{k=1}^K \widehat{\phi}_k)}{\prod_{k=1}^K \Gamma(\widehat{\phi}_k)} \right) - \log \left(\frac{\Gamma(K\phi)}{(\Gamma(\phi))^K} \right) - \frac{M}{2} \sum_{k=1}^K \log(\xi\widehat{\sigma}_k^2) - \frac{KM}{2} \\
&\quad + \sum_{n=1}^N \sum_{k=1}^K \bar{z}_k^{(n)} \log \bar{z}_k^{(n)} + \sum_{k=1}^K (\widehat{\phi}_k - \phi - \bar{N}_k) (\Psi(\widehat{\phi}_k) - \Psi(\sum_{k'=1}^K \widehat{\phi}_{k'})) \\
&\quad + \sum_{k=1}^K \frac{\xi (\|\widehat{\mu}_k - \mu_0\|^2 + M\widehat{\sigma}_k^2)}{2} + \sum_{k=1}^K \frac{\bar{N}_k (M \log(2\pi) + M\widehat{\sigma}_k^2)}{2} \\
&\quad + \sum_{k=1}^K \frac{\bar{N}_k \|\bar{\mathbf{x}}_k - \widehat{\mu}_k\|^2 + \sum_{n=1}^N \bar{z}_k^{(n)} \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}_k\|^2}{2}. \tag{4.22}
\end{aligned}$$

The prior hyperparameters, (ϕ, μ_0, ξ) , can be estimated by the EVB learning (Section 2.1.6). Computing the partial derivatives, we have

$$\frac{\partial F}{\partial \phi} = K(\Psi(\phi) - \Psi(K\phi)) - \sum_{k=1}^K (\Psi(\widehat{\phi}_k) - \Psi(\sum_{k'=1}^K \widehat{\phi}_{k'})), \tag{4.23}$$

$$\frac{\partial F}{\partial \mu_0} = \xi \sum_{k=1}^K (\mu_0 - \widehat{\mu}_k), \tag{4.24}$$

$$\frac{\partial F}{\partial \xi} = -\frac{M}{2} \left(\frac{K}{\xi} - \sum_{k=1}^K \left(\frac{\|\widehat{\mu}_k - \mu_0\|^2}{M} + \widehat{\sigma}_k^2 \right) \right). \tag{4.25}$$

The stationary conditions $\frac{\partial F}{\partial \mu_0} = \mathbf{0}$ and $\frac{\partial F}{\partial \xi} = 0$ yield the following update rules:

$$\mu_0 = \frac{1}{K} \sum_{k=1}^K \widehat{\mu}_k, \tag{4.26}$$

Algorithm 7 EVB learning for the Gaussian mixture model.

-
- 1: Initialize the variational parameters $(\{\widehat{\mathbf{z}}^{(n)}\}_{n=1}^N, \{\widehat{\phi}_k\}_{k=1}^K, \{\widehat{\mu}_k, \widehat{\sigma}_k^2\}_{k=1}^K)$, and the hyperparameters (ϕ, μ_0, ξ) .
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (4.21), (4.20), (4.12), (4.13), (4.16), (4.17), and (4.18) to update $\{\widehat{\mathbf{z}}^{(n)}\}_{n=1}^N, \{\widehat{\phi}_k\}_{k=1}^K$, and $\{\widehat{\mu}_k, \widehat{\sigma}_k^2\}_{k=1}^K$.
 - 3: Apply Eqs. (4.29), (4.26), and (4.27) to update ϕ, μ_0 and ξ , respectively.
 - 4: Evaluate the free energy (4.22).
 - 5: Iterate Steps 2 through 4 until convergence (until the energy decrease becomes smaller than a threshold).
-

$$\xi = \left\{ \frac{1}{K} \sum_{k=1}^K \left(\frac{\|\widehat{\mu}_k - \mu_0\|^2}{M} + \widehat{\sigma}_k^2 \right) \right\}^{-1}. \quad (4.27)$$

Since the stationary condition $\frac{\partial F}{\partial \phi} = 0$ is not explicitly solved for ϕ , the *Newton–Raphson* step is usually used for updating ϕ . With the second derivative,

$$\frac{\partial^2 F}{\partial \phi^2} = K \left(\Psi^{(1)}(\phi) - K \Psi^{(1)}(K\phi) \right), \quad (4.28)$$

the update rule is obtained as follows:

$$\begin{aligned} \phi^{\text{new}} &= \max \left(0, \phi^{\text{old}} - \left(\frac{\partial^2 F}{\partial \phi^2} \right)^{-1} \frac{\partial F}{\partial \phi} \right) \\ &= \max \left(0, \phi^{\text{old}} - \frac{K(\Psi(\phi) - \Psi(K\phi)) - \sum_{k=1}^K (\Psi(\widehat{\phi}_k) - \Psi(\sum_{k'=1}^K \widehat{\phi}_{k'})))}{K(\Psi^{(1)}(\phi) - K\Psi^{(1)}(K\phi))} \right), \end{aligned} \quad (4.29)$$

where $\Psi_m(z) \equiv \frac{d^m}{dz^m} \Psi(z)$ is the *polygamma function* of order m .

The EVB learning for the GMM is summarized in Algorithm 7. If the prior hyperparameters are fixed and Step 3 in the algorithm is omitted, the algorithm reduces to the (nonempirical) VB learning algorithm.

4.1.2 Mixture of Exponential Families

It is well known that the Gaussian distribution is an example of the *exponential family distribution*:

$$p(\mathbf{x}|\boldsymbol{\tau}) = p(\mathbf{t}|\boldsymbol{\eta}) = \exp \left(\boldsymbol{\eta}^\top \mathbf{t} - A(\boldsymbol{\eta}) + B(\mathbf{t}) \right), \quad (4.30)$$

where $\boldsymbol{\eta} \in \mathbf{H}$ is the *natural parameter*, $\boldsymbol{\eta}^\top \mathbf{t}$ is its inner product with the vector $\mathbf{t} = \mathbf{t}(\mathbf{x}) = (t_1(\mathbf{x}), \dots, t_M(\mathbf{x}))^\top$, and $A(\boldsymbol{\eta})$ and $B(\mathbf{t})$ are real-valued functions

of the parameter $\boldsymbol{\eta}$ and the sufficient statistics \mathbf{t} , respectively (Brown, 1986) (see Eq. (1.27) in Section 1.2.3). Suppose functions t_1, \dots, t_M and the constant function, 1, are linearly independent and the number of parameters in a single component distribution, $p(\mathbf{t}|\boldsymbol{\eta})$, is M .

The VB framework for GMMs in Section 4.1.1 is generalized to a mixture of exponential family distributions as follows. The conditional conjugate prior distributions of $\boldsymbol{\alpha} \in \Delta^{K-1}$ and $\{\boldsymbol{\eta}_k\}_{k=1}^K$ are given by

$$p(\boldsymbol{\alpha}|\phi) = \text{Dirichlet}_K(\boldsymbol{\alpha}; (\phi, \dots, \phi)^\top), \quad (4.31)$$

$$p(\{\boldsymbol{\eta}_k\}_{k=1}^K | \boldsymbol{\nu}_0, \xi) = \prod_{k=1}^K \frac{1}{C(\xi, \boldsymbol{\nu}_0)} \exp(\xi(\boldsymbol{\nu}_0^\top \boldsymbol{\eta}_k - A(\boldsymbol{\eta}_k))), \quad (4.32)$$

where the function $C(\xi, \boldsymbol{\nu})$ of $\xi \in \mathbb{R}$ and $\boldsymbol{\nu} \in \mathbb{R}^M$ is defined by

$$C(\xi, \boldsymbol{\nu}) = \int_H \exp(\xi(\boldsymbol{\nu}^\top \boldsymbol{\eta} - A(\boldsymbol{\eta}))) d\boldsymbol{\eta}. \quad (4.33)$$

Constants $\xi > 0$, $\boldsymbol{\nu}_0 \in \mathbb{R}^M$, and $\phi > 0$ are the hyperparameters.

VB Posterior for Mixture-of-Exponential-Family Models

Here, we derive the VB posterior $r_w(\mathbf{w})$ for the mixture-of-exponential-family model using Eq. (4.2).

Using the complete data $\{\mathbf{x}^{(n)}, \mathbf{z}^{(n)}\}_{n=1}^N$, we put

$$\bar{N}_k = \sum_{n=1}^N \langle z_k^{(n)} \rangle_{r_{\mathcal{H}}(\mathcal{H})}, \quad (4.34)$$

$$\bar{\mathbf{t}}_k = \frac{1}{\bar{N}_k} \sum_{n=1}^N \langle z_k^{(n)} \rangle_{r_{\mathcal{H}}(\mathcal{H})} \mathbf{t}^{(n)}, \quad (4.35)$$

where $\mathbf{t}^{(n)} = \mathbf{t}(\mathbf{x}^{(n)})$. Note that the variables \bar{N}_k and $\bar{\mathbf{t}}_k$ satisfy the constraints $\sum_{k=1}^K \bar{N}_k = N$ and $\sum_{k=1}^K \bar{N}_k \bar{\mathbf{t}}_k = \sum_{n=1}^N \mathbf{t}^{(n)}$. From Eq. (4.2) and the respective prior distributions, Eqs. (4.10) and (4.32), the VB posterior $r_w(\mathbf{w}) = r_\alpha(\boldsymbol{\alpha})r_\eta(\{\boldsymbol{\eta}_k\}_{k=1}^K)$ is obtained as the product of the following two distributions:

$$\begin{aligned} r_\alpha(\boldsymbol{\alpha}) &= \text{Dirichlet}_K(\boldsymbol{\alpha}; (\widehat{\phi}_1, \dots, \widehat{\phi}_K)^\top), \\ r_\eta(\{\boldsymbol{\eta}_k\}_{k=1}^K) &= \prod_{k=1}^K \frac{1}{C(\widehat{\xi}_k, \widehat{\boldsymbol{\nu}}_k)} \exp(\widehat{\xi}_k(\widehat{\boldsymbol{\nu}}_k^\top \boldsymbol{\eta}_k - A(\boldsymbol{\eta}_k))), \end{aligned} \quad (4.36)$$

where

$$\widehat{\phi}_k = \bar{N}_k + \phi, \quad (4.37)$$

$$\widehat{\boldsymbol{v}}_k = \frac{\overline{N}_k \bar{\boldsymbol{t}}_k + \xi \boldsymbol{v}_0}{\overline{N}_k + \xi}, \quad (4.38)$$

$$\widehat{\xi}_k = \overline{N}_k + \xi. \quad (4.39)$$

Let

$$\widehat{\boldsymbol{\eta}}_k = \langle \boldsymbol{\eta}_k \rangle_{r_\eta(\boldsymbol{\eta}_k)} = \frac{1}{\widehat{\xi}_k} \frac{\partial \log C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k)}{\partial \boldsymbol{v}_k}. \quad (4.40)$$

It follows that

$$\langle A(\boldsymbol{\eta}_k) \rangle_{r_\eta(\boldsymbol{\eta}_k)} = \widehat{\boldsymbol{\eta}}_k^\top \widehat{\boldsymbol{v}}_k - \frac{\partial \log C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k)}{\partial \xi_k}. \quad (4.41)$$

From Eq. (4.3), the VB posterior $r_{\mathcal{H}}(\mathcal{H})$ is given by

$$\begin{aligned} r_{\mathcal{H}}(\mathcal{H}) &= \prod_{n=1}^N r_z(z^{(n)}) \\ &= \prod_{n=1}^N \text{Multinomial}_{K,1}(z^{(n)}; \widehat{\boldsymbol{z}}^{(n)}), \end{aligned}$$

where $\widehat{\boldsymbol{z}}^{(n)} \in \Delta^{K-1}$ is

$$\widehat{z}_k^{(n)} = \langle z_k^{(n)} \rangle_{r_{\mathcal{H}}(\mathcal{H})} = \frac{\widehat{z}_k^{(n)}}{\sum_{k'=1}^K \widehat{z}_k^{(n)}}, \quad (4.42)$$

for

$$\widehat{z}_k^{(n)} = \exp \left(\Psi(\widehat{\phi}_k) - \Psi \left(\sum_{k'=1}^K \widehat{\phi}_{k'} \right) + \widehat{\boldsymbol{\eta}}_k^\top \boldsymbol{t}^{(n)} - \langle A(\boldsymbol{\eta}_k) \rangle_{r_\eta(\boldsymbol{\eta}_k)} + B(\boldsymbol{t}^{(n)}) \right). \quad (4.43)$$

To obtain the preceding expression of $\widehat{z}_k^{(n)}$, we used Eq. (4.19).

The free energy as a function of variational parameters is expressed as

$$\begin{aligned} F &= \log \left(\frac{\Gamma(\sum_{k=1}^K \widehat{\phi}_k)}{\prod_{k=1}^K \Gamma(\widehat{\phi}_k)} \right) - \log \left(\frac{\Gamma(K\phi)}{(\Gamma(\phi))^K} \right) - \sum_{k=1}^K \log C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k) + K \log C(\xi, \boldsymbol{v}_0) \\ &\quad + \sum_{n=1}^N \sum_{k=1}^K \widehat{z}_k^{(n)} \log \widehat{z}_k^{(n)} + \sum_{k=1}^K (\widehat{\phi}_k - \phi - \overline{N}_k) (\Psi(\widehat{\phi}_k) - \Psi(\sum_{k'=1}^K \widehat{\phi}_{k'})) \\ &\quad + \sum_{k=1}^K \left[\widehat{\boldsymbol{\eta}}_k^\top \left\{ \xi (\widehat{\boldsymbol{v}}_k - \boldsymbol{v}_0) + \overline{N}_k (\widehat{\boldsymbol{v}}_k - \bar{\boldsymbol{t}}_k) \right\} + (\widehat{\xi}_k - \xi - \overline{N}_k) \frac{\partial \log C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k)}{\partial \xi_k} \right] \\ &\quad - \sum_{n=1}^N B(\boldsymbol{t}^{(n)}). \end{aligned} \quad (4.44)$$

The update rule of ϕ for the EVB learning is obtained by Eq. (4.29) as in the GMM. The partial derivatives of F with respect to the hyperparameters (ν_0, ξ) are

$$\frac{\partial F}{\partial \nu_0} = K \frac{\partial \log C(\xi, \nu_0)}{\partial \nu_0} - \xi \sum_{k=1}^K \widehat{\boldsymbol{\eta}}_k, \quad (4.45)$$

$$\frac{\partial F}{\partial \xi} = \sum_{k=1}^K \left\{ \widehat{\boldsymbol{\eta}}_k^\top (\widehat{\nu}_k - \nu_0) - \frac{\partial \log C(\widehat{\xi}_k, \widehat{\nu}_k)}{\partial \xi_k} \right\} + K \frac{\partial \log C(\xi, \nu_0)}{\partial \xi}. \quad (4.46)$$

Equating these derivatives to zeros, we have the following stationary conditions:

$$\frac{1}{\xi} \frac{\partial \log C(\xi, \nu_0)}{\partial \nu_0} = \frac{1}{K} \sum_{k=1}^K \widehat{\boldsymbol{\eta}}_k, \quad (4.47)$$

$$\frac{\partial \log C(\xi, \nu_0)}{\partial \xi} = \left(\frac{1}{K} \sum_{k=1}^K \widehat{\boldsymbol{\eta}}_k \right)^\top \nu_0 - \frac{1}{K} \sum_{k=1}^K \langle A(\boldsymbol{\eta}_k) \rangle_{r_\eta(\boldsymbol{\eta}_k)}, \quad (4.48)$$

where we have used Eq. (4.41). If these equations are solved for ν_0 and ξ , respectively, we obtain their update rules as in the case of the GMM. Otherwise, we need the Newton–Raphson steps to update them.

The EVB learning for the mixture of exponential families is summarized in Algorithm 8. If the prior hyperparameters are fixed and Step 3 in the algorithm is omitted, the algorithm reduces to the (nonempirical) VB learning algorithm.

Algorithm 8 EVB learning for the mixture-of-exponential-family model.

- 1: Initialize the variational parameters $(\{\widehat{\mathbf{z}}^{(n)}\}_{n=1}^N, \{\widehat{\phi}_k\}_{k=1}^K, \{\widehat{\nu}_k, \widehat{\xi}_k\}_{k=1}^K)$, and the hyperparameters (ϕ, ν_0, ξ) .
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (4.43), (4.42), (4.34), (4.35), (4.37), (4.38), and (4.39) to update $\{\widehat{\mathbf{z}}^{(n)}\}_{n=1}^N, \{\widehat{\phi}_k\}_{k=1}^K$, and $\{\widehat{\nu}_k, \widehat{\xi}_k\}_{k=1}^K$. Transform $\{\widehat{\nu}_k, \widehat{\xi}_k\}_{k=1}^K$ to $\{\widehat{\boldsymbol{\eta}}_k, \langle A(\boldsymbol{\eta}_k) \rangle_{r_\eta(\boldsymbol{\eta}_k)}\}_{k=1}^K$ by Eqs. (4.40) and (4.41).
 - 3: Apply Eqs. (4.29), (4.47), and (4.48) to update ϕ, ν_0 and ξ , respectively.
 - 4: Evaluate the free energy (4.44).
 - 5: Iterate Steps 2 through 4 until convergence (until the energy decrease becomes smaller than a threshold).
-

4.1.3 Infinite Mixture Models

In 2000s, there was a revival of Bayesian nonparametric models to estimate the model complexity, e.g., the number of components in mixture models, by using a prior distribution for probability measures such as the *Dirichlet process (DP) prior*. The Bayesian nonparametric approach fits a single model adapting its complexity to the data. The VB framework plays an important role in achieving tractable inference for Bayesian nonparametric models. Here, we introduce the VB learning for the *stick-breaking* construction of the DP prior by instantiating the estimation of the number of components of the mixture model.

For the finite mixture model with K components, we had the discrete latent variable,

$$z \in \{e_1, e_2, \dots, e_K\},$$

indicating the label of the component. We also assumed the multinomial distribution,

$$p(z|\alpha) = \text{Multinomial}_{K,1}(z; \alpha) = \prod_{k=1}^K \alpha_k^{z_k}.$$

In the nonparametric Bayesian approach, we consider possibly an infinite number of components,

$$p(z|\alpha) = \lim_{K \rightarrow \infty} \text{Multinomial}_{K,1}(z; \alpha),$$

and the following generation process of α_k , called the stick-breaking process (Blei and Jordan, 2005; Gershman and Blei, 2012):

$$\begin{aligned} \alpha_k &= v_k \prod_{l=1}^{k-1} (1 - v_l), \\ v_k &\sim \text{Beta}(1, \gamma), \end{aligned}$$

where $\text{Beta}(\alpha, \beta)$ denotes the beta distribution with parameters α and β , and γ is the scaling parameter.

To derive a tractable VB learning algorithm, the truncation level T is usually introduced to the preceding process, which enforces $v_T = 1$. If the truncation level T is sufficiently large, some components are left unused, and hence T does not directly specify the number of components.

Then, the VB posterior $r(\mathcal{H}, v)$ for the latent variables and $v = \{v_k\}_{k=1}^{T-1}$ is assumed to be factorized, $r_{\mathcal{H}}(\mathcal{H})r_v(v)$, for which the free energy minimization implies further factorization:

$$r(\mathcal{H}, v) = \prod_{n=1}^N r_z(z^{(n)}) \prod_{k=1}^{T-1} r_v(v_k),$$

where $r_z(z^{(n)})$ is the multinomial distribution as in the case of the finite mixture model, and $r_v(v_k)$ is the beta distribution because of the conditional conjugacy. To see this and how the VB learning algorithm is derived, we instantiate the GMM discussed in Section 4.1.1.

The free energy is decomposed as

$$\begin{aligned} F = & \left\langle \log \frac{r_z(\{z^{(n)}\}_{n=1}^N) r_v(v) r_\mu(\{\mu_k\}_{k=1}^K)}{p(v)p(\{\mu_k\}_{k=1}^T)} \right\rangle_{r_z(\{z^{(n)}\}_{n=1}^N) r_v(v) r_\mu(\{\mu_k\}_{k=1}^K)} \\ & - \left\langle \log p(\mathcal{D} | \{z^{(n)}\}_{n=1}^N, w) \right\rangle_{r_z(\{z^{(n)}\}_{n=1}^N) r_v(v) r_\mu(\{\mu_k\}_{k=1}^K)} \\ & - \left\langle \log p(\{z^{(n)}\}_{n=1}^N | v) \right\rangle_{r_z(\{z^{(n)}\}_{n=1}^N) r_v(v)}. \end{aligned}$$

These terms are computed in the same way as in Section 4.1.1 except for the last term, $\left\langle \log p(\{z^{(n)}\}_{n=1}^N | v) \right\rangle_{r_z(\{z^{(n)}\}_{n=1}^N) r_v(v)} = \sum_{n=1}^N \left\langle \log p(z^{(n)} | v) \right\rangle_{r_z(z^{(n)}) r_v(v)}$.

Let $c^{(n)}$ be the index k such that $z_k^{(n)} = 1$ and θ be the indicator function. Then, we have

$$\begin{aligned} & \left\langle \log p(z^{(n)} | v) \right\rangle_{r_z(z^{(n)}) r_v(v)} \\ &= \left\langle \log \prod_{k=1}^{\infty} (1 - v_k)^{\theta(c^{(n)} > k)} v_k^{\theta(c^{(n)} = k)} \right\rangle_{r_z(z^{(n)}) r_v(v)} \\ &= \sum_{k=1}^{\infty} \left\{ r_z(c^{(n)} > k) \langle \log(1 - v_k) \rangle_{r_v(v)} + r_z(c^{(n)} = k) \langle \log v_k \rangle_{r_v(v)} \right\} \\ &= \sum_{k=1}^{T-1} \left\{ r_z(c^{(n)} > k) \langle \log(1 - v_k) \rangle_{r_v(v)} + r_z(c^{(n)} = k) \langle \log v_k \rangle_{r_v(v)} \right\}, \end{aligned}$$

where we have used $\log v_T = 0$ and $r_z(c^{(n)} > T) = 0$.

Since the probabilities $r_z(c^{(n)} = k)$ and $r_z(c^{(n)} > k)$ are given by

$$r_z(c^{(n)} = k) = \bar{z}_k^{(n)},$$

$$r_z(c^{(n)} > k) = \sum_{l=k+1}^T \bar{z}_l^{(n)},$$

it follows from Eq. (4.12) that

$$\begin{aligned} \sum_{n=1}^N r_z(c^{(n)} = k) &= \bar{N}_k, \\ \sum_{n=1}^N r_z(c^{(n)} > k) &= \sum_{l=k+1}^T \bar{N}_l = N - \sum_{l=1}^k \bar{N}_l. \end{aligned}$$

Now Eq. (4.3) in this case yields that

$$r_v(\mathbf{v}) \propto \sum_{n=1}^N \left\langle \log p(z^{(n)}|\mathbf{v}) \right\rangle_{r_z(z^{(n)})} p(\mathbf{v}).$$

It follows from similar manipulations to those just mentioned and the conditional conjugacy that

$$r_v(\mathbf{v}) = \prod_{k=1}^{T-1} \text{Beta}(v_k; \widehat{\kappa}_k, \widehat{\lambda}_k), \quad (4.49)$$

i.e., for a fixed $r_{\mathcal{H}}(\mathcal{H})$, the optimal $r_v(\mathbf{v})$ is the beta distribution with the parameters,

$$\widehat{\kappa}_k = 1 + \bar{N}_k, \quad (4.50)$$

$$\widehat{\lambda}_k = \gamma + N - \sum_{l=1}^k \bar{N}_l. \quad (4.51)$$

The VB posterior $r_{\mathcal{H}}(\mathcal{H})$ is computed similarly except that the expectation

$$\langle \log \alpha_k \rangle_{r_a(\alpha)} = \Psi(\bar{N}_k + \phi) - \Psi(N + K\phi)$$

in Eq. (4.19) for the finite mixture model is replaced by

$$\langle \log v_k \rangle_{r_v(v_k)} + \sum_{l=1}^{k-1} \langle \log(1 - v_l) \rangle_{r_v(v_l)} = \Psi(\widehat{\kappa}_k) - \Psi(\widehat{\kappa}_k + \widehat{\lambda}_k) + \sum_{l=1}^{k-1} \{\Psi(\widehat{\lambda}_l) - \Psi(\widehat{\kappa}_l + \widehat{\lambda}_l)\},$$

since

$$\begin{aligned} \langle \log v_k \rangle_{r_v(v_k)} &= \Psi(\widehat{\kappa}_k) - \Psi(\widehat{\kappa}_k + \widehat{\lambda}_k), \\ \langle \log(1 - v_k) \rangle_{r_v(v_k)} &= \Psi(\widehat{\lambda}_k) - \Psi(\widehat{\kappa}_k + \widehat{\lambda}_k), \end{aligned}$$

for $r_v(v_k) = \text{Beta}(v_k; \widehat{\kappa}_k, \widehat{\lambda}_k)$. In the case of the GMM, $\bar{z}_k^{(n)}$ in Eq. (4.21) is replaced with

$$\begin{aligned} \bar{z}_k^{(n)} &= \exp \left(\Psi(\widehat{\kappa}_k) - \Psi(\widehat{\kappa}_k + \widehat{\lambda}_k) + \sum_{l=1}^{k-1} \{\Psi(\widehat{\lambda}_l) - \Psi(\widehat{\kappa}_l + \widehat{\lambda}_l)\} \right. \\ &\quad \left. - \frac{1}{2} \|\mathbf{x}^{(n)} - \widehat{\mu}_k\|^2 + M\widehat{\sigma}_k^2 \right). \end{aligned} \quad (4.52)$$

The free energy is given by

$$\begin{aligned} F &= \sum_{k=1}^{T-1} \log \left(\frac{\Gamma(\widehat{\kappa}_k + \widehat{\lambda}_k)}{\Gamma(\widehat{\kappa}_k)\Gamma(\widehat{\lambda}_k)} \right) - (T-1) \log \gamma - \frac{M}{2} \sum_{k=1}^T \log (\xi \widehat{\sigma}_k^2) - \frac{TM}{2} \\ &\quad + \sum_{n=1}^N \sum_{k=1}^T \widehat{z}_k^{(n)} \log \widehat{z}_k^{(n)} + \sum_{k=1}^{T-1} (\widehat{\kappa}_k - 1 - \bar{N}_k) \{\Psi(\widehat{\kappa}_k) - \Psi(\widehat{\kappa}_k + \widehat{\lambda}_k)\} \end{aligned}$$

$$\begin{aligned}
& + \sum_{k=1}^{T-1} \left\{ \widehat{\lambda}_k - \gamma - \left(N - \sum_{l=1}^k \bar{N}_l \right) \right\} \{ \Psi(\widehat{\lambda}_k) - \Psi(\widehat{\kappa}_k + \widehat{\lambda}_k) \} \\
& + \sum_{k=1}^T \frac{\xi (\|\widehat{\mu}_k - \mu_0\|^2 + M\widehat{\sigma}_k^2)}{2} + \sum_{k=1}^T \frac{\bar{N}_k (M \log(2\pi) + M\widehat{\sigma}_k^2)}{2} \\
& + \sum_{k=1}^T \frac{\bar{N}_k \|\bar{x}_k - \widehat{\mu}_k\|^2 + \sum_{n=1}^N \widehat{z}_k^{(n)} \|\mathbf{x}^{(n)} - \bar{x}_k\|^2}{2}. \tag{4.53}
\end{aligned}$$

The VB learning algorithm is similar to Algorithm 7 for the finite GMM while the number of components K is replaced with the truncation level T throughout, the update rule (4.21) is replaced with Eq. (4.52), and $\{\widehat{\kappa}_k, \widehat{\lambda}_k\}_{k=1}^{T-1}$ are updated by Eqs. (4.50) and (4.51) instead of $\{\widehat{\phi}_k\}_{k=1}^K$.

By computing $\frac{\partial F}{\partial \gamma}$ and equating it to zero, the EVB learning for the hyperparameter γ updates it as follows:

$$\gamma = \left[\frac{-1}{T-1} \sum_{k=1}^{T-1} \{ \Psi(\widehat{\lambda}_k) - \Psi(\widehat{\kappa}_k + \widehat{\lambda}_k) \} \right]^{-1}, \tag{4.54}$$

which can replace the update rule of ϕ in Step 3 of Algorithm 7.

4.2 Other Latent Variable Models

In this section, we discuss more complex latent variable models than mixture models and derive VB learning algorithms for them. Although we focus on the models where the multinomial distribution is assumed on the observed data given latent variables, it is straightforward to replace it with other members of the exponential family.

4.2.1 Bayesian Networks

A *Bayesian network* is a probabilistic model defined by a graphical model expressing the relations among random variables by a graph and the conditional probabilities associated with them (Jensen, 2001). In this subsection, we focus on a Bayesian network whose states of all hidden nodes influence those of all observation nodes, and assume that it has M observation nodes and K hidden nodes. The graphical structure of this Bayesian network is called bipartite and presented in Figure 4.1.

The observation nodes are denoted by $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_M)$, and the set of states of observation node $\mathbf{x}_j = (x_{j,1}, \dots, x_{j,Y_j})^\top \in \{\mathbf{e}_l\}_{l=1}^{Y_j}$ is $\{1, \dots, Y_j\}$. The hidden

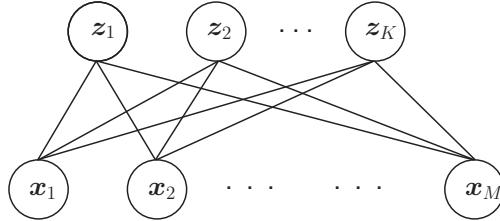


Figure 4.1 Graphical structure of the Bayesian network.

nodes are denoted by $\mathbf{z} = (z_1, \dots, z_K)$, and the set of states of hidden node $z_k = (z_{k,1}, \dots, z_{k,T_k})^\top \in \{\mathbf{e}_i\}_{i=1}^{T_k}$ is $\{1, \dots, T_k\}$.

The probability that the state of the hidden node z_k is i ($1 \leq i \leq T_k$) is expressed as

$$a_{(k,i)} = \text{Prob}(z_k = \mathbf{e}_i).$$

Then, $\mathbf{a}_k = (a_{(k,1)}, \dots, a_{(k,T_k)})^\top \in \Delta^{T_k-1}$ for $k = 1, \dots, K$.

The conditional probability that the j th observation node x_j is l ($1 \leq l \leq Y_j$), given the condition that the states of hidden nodes are $\mathbf{z} = (z_1, \dots, z_K)$, is denoted by

$$b_{(j,l|\mathbf{z})} = \text{Prob}(x_j = \mathbf{e}_l | \mathbf{z}).$$

Then, $\mathbf{b}_{j|\mathbf{z}} = (b_{(j,1|\mathbf{z})}, \dots, b_{(j,Y_j|\mathbf{z})})^\top \in \Delta^{Y_j-1}$ for $j = 1, \dots, M$. Define $\mathbf{b}_z = \{\mathbf{b}_{j|\mathbf{z}}\}_{j=1}^M$ for $\mathbf{z} \in \mathcal{Z} = \{\mathbf{z}; z_k \in \{\mathbf{e}_i\}_{i=1}^{T_k}, k = 1, \dots, K\}$. Let $\mathbf{w} = \{\{\mathbf{a}_k\}_{k=1}^K, \{\mathbf{b}_z\}_{z \in \mathcal{Z}}\}$ be the set of all parameters. Then, the joint probability that the states of observation nodes are $\mathbf{x} = (x_1, \dots, x_M)$ and the states of hidden nodes are $\mathbf{z} = (z_1, \dots, z_K)$ is

$$p(\mathbf{x}, \mathbf{z} | \mathbf{w}) = p(\mathbf{x} | \mathbf{b}_z) \prod_{k=1}^K \prod_{i=1}^{T_k} a_{(k,i)}^{z_{k,i}},$$

where

$$p(\mathbf{x} | \mathbf{b}_z) = \prod_{j=1}^M \prod_{l=1}^{Y_j} b_{(j,l|\mathbf{z})}^{x_{j,l}}.$$

Therefore, the marginal probability that the states of observation nodes are \mathbf{x} is

$$\begin{aligned} p(\mathbf{x} | \mathbf{w}) &= \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x}, \mathbf{z} | \mathbf{w}) \\ &= \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x} | \mathbf{b}_z) \prod_{k=1}^K \prod_{i=1}^{T_k} a_{(k,i)}^{z_{k,i}}, \end{aligned} \tag{4.55}$$

where we used the notation $\sum_{z \in \mathcal{Z}}$ for the summation over all states of hidden nodes. Let

$$M_{\text{obs}} = \sum_{j=1}^M (Y_j - 1),$$

which is the number of parameters to specify the conditional probability $p(\mathbf{x}|\mathbf{b}_z)$ of the states of all the observation nodes given the states of the hidden nodes. Then, the number of the parameters of the model, D , is

$$D = M_{\text{obs}} \prod_{k=1}^K T_k + \sum_{k=1}^K (T_k - 1). \quad (4.56)$$

We assume that the prior distribution $p(\mathbf{w})$ of the parameters $\mathbf{w} = \{\{\mathbf{a}_k\}_{k=1}^K, \{\mathbf{b}_z\}_{z \in \mathcal{Z}}\}$ is the conditional conjugate prior distribution. Then, $p(\mathbf{w})$ is given by $\left\{ \prod_{k=1}^K p(\mathbf{a}_k|\phi) \right\} \left\{ \prod_{z \in \mathcal{Z}} \prod_{j=1}^M p(\mathbf{b}_{j|z}|\xi) \right\}$, where

$$p(\mathbf{a}_k|\phi) = \text{Dirichlet}_{T_k}(\mathbf{a}_k; (\phi, \dots, \phi)^\top), \quad (4.57)$$

$$p(\mathbf{b}_{j|z}|\xi) = \text{Dirichlet}_{Y_j}(\mathbf{b}_{j|z}; (\xi, \dots, \xi)^\top), \quad (4.58)$$

are the Dirichlet distributions with hyperparameters $\phi > 0$ and $\xi > 0$.

VB Posterior for Bayesian Networks

Let $\{\mathcal{D}, \mathcal{H}\}$ be the complete data with the observed data set $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^N$ and the corresponding hidden variables $\mathcal{H} = \{\mathbf{z}^{(n)}\}_{n=1}^N$. Define the expected sufficient statistics:

$$\begin{aligned} \overline{N}_{(k,i_k)} &= \sum_{n=1}^N \left\langle z_{k,i_k}^{(n)} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})}, \\ \overline{N}_{(j,l_j|\mathbf{z})} &= \sum_{n=1}^N x_{j,l_j}^{(n)} r_z(\mathbf{z}^{(n)} = \mathbf{z}), \end{aligned} \quad (4.59)$$

where

$$r_z(\mathbf{z}^{(n)} = \mathbf{z}) = \left\langle \prod_{k=1}^K z_{k,i_k}^{(n)} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})} \quad (4.60)$$

is the estimated probability that $\mathbf{z}^{(n)} = \mathbf{z} = (\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_K})$. Here $x_j^{(n)}$ indicates the state of the j th observation node and $z_k^{(n)}$ indicates the state of the k th hidden node when the n th training sample is observed. From Eq. (4.2), the VB posterior distribution of parameters $\mathbf{w} = \{\{\mathbf{a}_k\}_{k=1}^K, \{\mathbf{b}_z\}_{z \in \mathcal{Z}}\}$ is given by

$$r_w(\mathbf{w}) = \left\{ \prod_{k=1}^K r_a(\mathbf{a}_k) \right\} \left\{ \prod_{z \in \mathcal{Z}} \prod_{j=1}^M r_b(\mathbf{b}_{j|z}) \right\},$$

$$r_a(\mathbf{a}_k) = \text{Dirichlet}_{T_k}(\mathbf{a}_k; \widehat{\boldsymbol{\phi}}_k), \quad (4.61)$$

$$r_b(\mathbf{b}_{j|z}) = \text{Dirichlet}_{Y_j}(\mathbf{b}_{j|z}; \widehat{\boldsymbol{\xi}}_{j|z}), \quad (4.62)$$

where

$$\widehat{\boldsymbol{\phi}}_k = (\widehat{\phi}_{(k,1)}, \dots, \widehat{\phi}_{(k,T_k)})^\top \quad (k = 1, \dots, K),$$

$$\widehat{\phi}_{(k,i)} = \overline{N}_{(k,i)}^z + \phi \quad (i = 1, \dots, T_k), \quad (4.63)$$

$$\widehat{\boldsymbol{\xi}}_{j|z} = (\widehat{\xi}_{(j,1|z)}, \dots, \widehat{\xi}_{(j,Y_j|z)})^\top \quad (j = 1, \dots, M, z \in \mathcal{Z}),$$

$$\widehat{\xi}_{(j,l|z)} = \overline{N}_{(j,l|z)}^x + \xi \quad (l = 1, \dots, Y_j). \quad (4.64)$$

Note that if we define

$$\overline{N}_z^x = \sum_{n=1}^N r_z(z^{(n)} = z) = \sum_{n=1}^N \left\langle \prod_{k=1}^K z_{k,i_k}^{(n)} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})},$$

for $\mathbf{z} = (\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_K}) \in \mathcal{Z}$, we have

$$\overline{N}_z^x = \sum_{l=1}^{Y_j} \overline{N}_{(j,l|z)}^x, \quad (4.65)$$

for $j = 1, \dots, M$, and

$$\overline{N}_{(k,i)}^z = \sum_{z_{-k}} \overline{N}_z^x, \quad (4.66)$$

where $\sum_{z_{-k}}$ denotes the summation over $z_{k'}$ ($k' \neq k$) other than $z_k = \mathbf{e}_i$.

It follows from Eqs. (4.61) and (4.62) that

$$\langle \log a_{(k,i)} \rangle_{r_a(\mathbf{a}_k)} = \Psi(\widehat{\phi}_{(k,i)}) - \Psi\left(\sum_{i'=1}^{T_k} \widehat{\phi}_{(k,i')}\right) \quad (i = 1, \dots, T_k),$$

for $k = 1, \dots, K$ and

$$\langle \log b_{(j,l|z)} \rangle_{r_b(\mathbf{b}_{j|z})} = \Psi(\widehat{\xi}_{(j,l|z)}) - \Psi\left(\sum_{l'=1}^{Y_j} \widehat{\xi}_{(j,l'|z)}\right) \quad (l = 1, \dots, Y_j),$$

for $j = 1, \dots, M$. From Eq. (4.3), the VB posterior distribution of the hidden variables is given by $r_{\mathcal{H}}(\mathcal{H}) = \prod_{n=1}^N r_z(z^{(n)})$, where for $\mathbf{z} = (\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_K})$,

$$\begin{aligned} r_z(z^{(n)} = z) &= \sum_{z^{(n)} \in \mathcal{Z}} r_z(z^{(n)}) \prod_{k=1}^K z_{k,i_k}^{(n)} \\ &\propto \exp \left(\sum_{k=1}^K \left\{ \Psi(\widehat{\phi}_{(k,i_k)}) - \Psi\left(\sum_{i'_k=1}^{T_k} \widehat{\phi}_{(k,i'_k)}\right) \right\} \right. \\ &\quad \left. + \sum_{j=1}^M \left\{ \Psi(\widehat{\xi}_{(j,l^{(n)}|z)}) - \Psi\left(\sum_{l'=1}^{Y_j} \widehat{\xi}_{(j,l'|z)}\right) \right\} \right), \end{aligned} \quad (4.67)$$

if $x_j^{(n)} = \mathbf{e}_{i_j^{(n)}}$.

The VB algorithm updates $\{\bar{N}_{(j,l|z)}^x\}$ using Eqs. (4.59) and (4.67) iteratively. The other expected sufficient statistics and variational parameters are computed by Eqs. (4.65), (4.66) and Eqs. (4.63), (4.64), respectively. The free energy as a function of the variational parameters is given by

$$\begin{aligned} F = & \sum_{k=1}^K \left\{ \log \left(\frac{\Gamma(\sum_{i=1}^{T_k} \widehat{\phi}_{(k,i)})}{\prod_{i=1}^{T_k} \Gamma(\widehat{\phi}_{(k,i)})} \right) - \log \left(\frac{\Gamma(T_k \phi)}{(\Gamma(\phi))^{T_k}} \right) \right. \\ & + \sum_{i=1}^{T_k} (\widehat{\phi}_{(k,i)} - \phi - \bar{N}_{(k,i)}^z) (\Psi(\widehat{\phi}_{(k,i)}) - \Psi(\sum_{i'=1}^{T_k} \widehat{\phi}_{(k,i')})) \Big\} \\ & + \sum_{z \in \mathcal{Z}} \sum_{j=1}^M \left\{ \log \left(\frac{\Gamma(\sum_{l=1}^{Y_j} \widehat{\xi}_{(j,l|z)})}{\prod_{l=1}^{Y_j} \Gamma(\widehat{\xi}_{(j,l|z)})} \right) - \log \left(\frac{\Gamma(Y_j \xi)}{(\Gamma(\xi))^{Y_j}} \right) \right. \\ & + \sum_{l=1}^{Y_j} (\widehat{\xi}_{(j,l|z)} - \xi - \bar{N}_{(j,l|z)}^x) (\Psi(\widehat{\xi}_{(j,l|z)}) - \Psi(\sum_{l'=1}^{Y_j} \widehat{\xi}_{(j,l'|z)})) \Big\} \\ & + \sum_{n=1}^N \sum_{z \in \mathcal{Z}} r_z(z^{(n)} = z) \log r_z(z^{(n)} = z). \end{aligned} \quad (4.68)$$

The following update rule for the EVB learning of the hyperparameter ϕ is obtained in the same way as the update rule (4.29) for the GMM:

$$\phi^{\text{new}} = \max \left(0, \phi^{\text{old}} - \frac{\sum_{k=1}^K \{ T_k (\Psi(\phi) - \Psi(T_k \phi)) - \sum_{i=1}^{T_k} (\Psi(\widehat{\phi}_{(k,i)}) - \Psi(\sum_{i'=1}^{T_k} \widehat{\phi}_{(k,i')})) \}}{\sum_{k=1}^K T_k (\Psi^{(1)}(\phi) - T_k \Psi^{(1)}(T_k \phi))} \right). \quad (4.69)$$

Similarly, we obtain the following update rule of the hyperparameter ξ :

$$\xi^{\text{new}} = \max \left(0, \xi^{\text{old}} - \frac{\sum_{z \in \mathcal{Z}} \sum_{j=1}^M \{ Y_j (\Psi(\xi) - \Psi(Y_j \xi)) - \sum_{l=1}^{Y_j} (\Psi(\widehat{\xi}_{(j,l|z)}) - \Psi(\sum_{l'=1}^{Y_j} \widehat{\xi}_{(j,l'|z)})) \}}{(\prod_{k=1}^K T_k) \sum_{j=1}^M Y_j (\Psi^{(1)}(\xi) - Y_j \Psi^{(1)}(Y_j \xi))} \right). \quad (4.70)$$

Let $\widehat{\mathcal{S}} = \{r_z(z^{(n)} = z)\}_{n=1,z \in \mathcal{Z}}^N = \left\{ \left\langle \prod_{k=1}^K z_{k,i_k}^{(n)} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})} \right\}_{n=1,z \in \mathcal{Z}}^N$, $\widehat{\Phi} = \{\widehat{\phi}_k\}_{k=1}^K$, and $\widehat{\Xi} = \{\widehat{\xi}_{j|z}\}_{j=1,z \in \mathcal{Z}}^M$ be the sets of variational parameters. The EVB learning for the Bayesian network is summarized in Algorithm 9. If the prior hyperparameters are fixed and Step 3 in the algorithm is omitted, the algorithm reduces to the (nonempirical) VB learning algorithm.

4.2.2 Hidden Markov Models

Hidden Markov models (HMMs) have been widely used for sequence modeling in speech recognition, natural language processing, and so on (Rabiner,

Algorithm 9 EVB learning for the Bayesian network.

-
- 1: Initialize the variational parameters $(\widehat{\mathcal{S}}, \widehat{\Phi}, \widehat{\Xi})$ and the hyperparameters (ϕ, ξ) .
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (4.67), (4.59), (4.65), (4.66), (4.63), and (4.64) to update $\widehat{\mathcal{S}}, \widehat{\Phi}$, and $\widehat{\Xi}$.
 - 3: Apply Eqs. (4.69) and (4.70) to update ϕ and ξ , respectively.
 - 4: Evaluate the free energy (4.68).
 - 5: Iterate Steps 2 through 4 until convergence (until the energy decrease becomes smaller than a threshold).
-

1989). In this subsection, we consider discrete HMMs. Suppose a sequence $\mathcal{D} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)})$ was observed. Each $\mathbf{x}^{(t)}$ is an M -dimensional binary vector (M -valued finite alphabet):

$$\mathbf{x}^{(t)} = (x_1^{(t)}, \dots, x_M^{(t)}) \in \{\mathbf{e}_1, \dots, \mathbf{e}_M\},$$

where if the output symbol at time t is m , then $x_m^{(t)} = 1$, and otherwise 0. Moreover, $\mathbf{x}^{(t)}$ is produced in K -valued discrete hidden state $\mathbf{z}^{(t)}$. The sequence of hidden states $\mathcal{H} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(T)})$ is generated by a first-order Markov process. Similarly, $\mathbf{z}^{(t)}$ is represented by a K -dimensional binary vector

$$\mathbf{z}^{(t)} = (z_1^{(t)}, \dots, z_K^{(t)}) \in \{\mathbf{e}_1, \dots, \mathbf{e}_K\},$$

where if the hidden state at time t is k , then $z_k^{(t)} = 1$, and otherwise 0.

Without loss of generality, we assume that the initial state ($t = 1$) is the first one, namely $z_1^{(1)} = 1$. Then, the probability of a sequence is given by

$$p(\mathcal{D}|\mathbf{w}) = \sum_{\mathcal{H}} \prod_{m=1}^M b_{1,m}^{x_m^{(1)}} \prod_{t=2}^T \prod_{k=1}^K \prod_{l=1}^K a_{k,l}^{z_l^{(t)} z_k^{(t-1)}} \prod_{m=1}^M b_{k,m}^{z_k^{(t)} x_m^{(t)}}, \quad (4.71)$$

where $\sum_{\mathcal{H}}$ is taken all over possible values of hidden variables, and the model parameters, $\mathbf{w} = (\mathbf{A}, \mathbf{B})$, consist of the state transition probability matrix $\mathbf{A} = (\widetilde{\mathbf{a}}_1, \dots, \widetilde{\mathbf{a}}_K)^{\top}$ and the emission probability matrix $\mathbf{B} = (\widetilde{\mathbf{b}}_1, \dots, \widetilde{\mathbf{b}}_K)^{\top}$ satisfying $\widetilde{\mathbf{a}}_k = (a_{k,1}, \dots, a_{k,K})^{\top} \in \Delta^{K-1}$ and $\widetilde{\mathbf{b}}_k = (b_{k,1}, \dots, b_{k,K})^{\top} \in \Delta^{M-1}$ for $1 \leq k \leq K$, respectively. $a_{k,l}$ represents the transition probability from the k th hidden state to the l th hidden state and $b_{k,m}$ is the emission probability that alphabet m is produced in the k th hidden state. Figure 4.2 illustrates an example of the state transition diagram of an HMM.

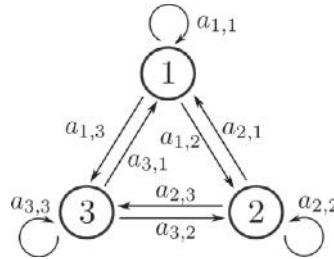


Figure 4.2 State transition diagram of an HMM.

The log-likelihood of the HMM for a sequence of complete data $\{\mathcal{D}, \mathcal{H}\}$ is defined by

$$\log p(\mathcal{D}, \mathcal{H} | \mathbf{w}) = \sum_{t=2}^T \sum_{k=1}^K \sum_{l=1}^K z_k^{(t)} z_l^{(t-1)} \log a_{k,l} + \sum_{t=1}^T \sum_{k=1}^K \sum_{m=1}^M z_k^{(t)} x_m^{(t)} \log b_{k,m}.$$

We assume that the prior distributions of the transition probability matrix \mathbf{A} and the emission probability matrix \mathbf{B} are the Dirichlet distributions with hyperparameters $\phi > 0$ and $\xi > 0$:

$$p(\mathbf{A} | \phi) = \prod_{k=1}^K \text{Dirichlet}_K \left(\tilde{\mathbf{a}}_k; (\phi, \dots, \phi)^\top \right), \quad (4.72)$$

$$p(\mathbf{B} | \xi) = \prod_{k=1}^K \text{Dirichlet}_M \left(\tilde{\mathbf{b}}_k; (\xi, \dots, \xi)^\top \right). \quad (4.73)$$

VB Posterior for HMMs

We define the expected sufficient statistics by

$$\bar{N}_k = \sum_{t=1}^T \langle z_k^{(t)} \rangle_{r_{\mathcal{H}}(\mathcal{H})}, \quad (4.74)$$

$$\bar{N}_{k,l}^{[z]} = \sum_{t=2}^T \langle z_l^{(t)} z_k^{(t-1)} \rangle_{r_{\mathcal{H}}(\mathcal{H})}, \quad (4.75)$$

$$\bar{N}_{k,m}^{[x]} = \sum_{t=1}^T \langle z_k^{(t)} \rangle_{r_{\mathcal{H}}(\mathcal{H})} x_m^{(t)}, \quad (4.76)$$

where the expected count \bar{N}_k is constrained by $\bar{N}_k = \sum_l \bar{N}_{k,l}^{[z]}$. Then, the VB posterior distribution of parameters $r_w(\mathbf{w})$ is given by

$$r_A(\mathbf{A}) = \prod_{k=1}^K \text{Dirichlet}_K \left(\tilde{\mathbf{a}}_k; (\widehat{\phi}_{k,1}, \dots, \widehat{\phi}_{k,K})^\top \right),$$

$$r_B(\mathbf{B}) = \prod_{k=1}^K \text{Dirichlet}_M \left(\tilde{\mathbf{b}}_k; (\widehat{\xi}_{k,1}, \dots, \widehat{\xi}_{k,M})^\top \right),$$

where

$$\widehat{\phi}_{k,l} = \overline{N}_{k,l}^{[z]} + \phi, \quad (4.77)$$

$$\widehat{\xi}_{k,m} = \overline{N}_{k,m}^{[x]} + \xi. \quad (4.78)$$

The posterior distribution of hidden variables $r_{\mathcal{H}}(\mathcal{H})$ is given by

$$r_{\mathcal{H}}(\mathcal{H}) = \frac{1}{C_{\mathcal{H}}} \exp \left(\sum_{t=2}^T \sum_{k=1}^K \sum_{l=1}^K z_k^{(t)} z_l^{(t-1)} \langle \log a_{k,l} \rangle_{r_A(\mathbf{A})} \right. \\ \left. + \sum_{t=1}^T \sum_{k=1}^K \sum_{m=1}^M z_k^{(t)} x_m^{(t)} \langle \log b_{k,m} \rangle_{r_B(\mathbf{B})} \right), \quad (4.79)$$

where $C_{\mathcal{H}}$ is the normalizing constant and

$$\langle \log a_{k,l} \rangle_{r_A(\mathbf{A})} = \Psi(\widehat{\phi}_{k,l}) - \Psi \left(\sum_{l'=1}^K \widehat{\phi}_{k,l'} \right),$$

$$\langle \log b_{k,m} \rangle_{r_B(\mathbf{B})} = \Psi(\widehat{\xi}_{k,m}) - \Psi \left(\sum_{m'=1}^M \widehat{\xi}_{k,m'} \right).$$

The expected sufficient statistics $\langle z_k^{(t)} \rangle_{r_{\mathcal{H}}(\mathcal{H})}$ and $\langle z_l^{(t)} z_k^{(t-1)} \rangle_{r_{\mathcal{H}}(\mathcal{H})}$ in Eqs. (4.74) through (4.76) can be efficiently computed in the order of $O(T)$ by the *forward–backward algorithm* (Beal, 2003). This algorithm can also compute $C_{\mathcal{H}}$. Thus, after the substitution of Eq. (4.3), the free energy is given by

$$F = \sum_{k=1}^K \left\{ \log \left(\frac{\Gamma(\sum_{l=1}^K \widehat{\phi}_{k,l})}{\prod_{l=1}^K \Gamma(\widehat{\phi}_{k,l})} \right) + \sum_{l=1}^K (\widehat{\phi}_{k,l} - \phi) (\Psi(\widehat{\phi}_{k,l}) - \Psi(\sum_{l'=1}^K \widehat{\phi}_{k,l'})) \right. \\ \left. + \log \left(\frac{\Gamma(\sum_{m=1}^M \widehat{\xi}_{k,m})}{\prod_{m=1}^M \Gamma(\widehat{\xi}_{k,m})} \right) + \sum_{m=1}^M (\widehat{\xi}_{k,m} - \xi) (\Psi(\widehat{\xi}_{k,m}) - \Psi(\sum_{m'=1}^M \widehat{\xi}_{k,m'})) \right\} \\ - K \log \left(\frac{\Gamma(K\phi)}{(\Gamma(\phi))^K} \right) - K \log \left(\frac{\Gamma(M\xi)}{(\Gamma(\xi))^M} \right) - \log C_{\mathcal{H}}. \quad (4.80)$$

The following update rule for the EVB learning of the hyperparameter ϕ is obtained in the same way as the update rule (4.29) for the GMM:

$$\phi^{\text{new}} = \max \left(0, \phi^{\text{old}} - \frac{K^2 (\Psi(\phi) - \Psi(K\phi)) - \sum_{k=1}^K \sum_{l=1}^K (\Psi(\widehat{\phi}_{k,l}) - \Psi(\sum_{l'=1}^K \widehat{\phi}_{k,l'}))}{K^2 (\Psi^{(1)}(\phi) - K\Psi^{(1)}(K\phi))} \right). \quad (4.81)$$

Algorithm 10 EVB learning for the hidden Markov model.

-
- 1: Initialize the variational parameters $(\widehat{\mathcal{S}}, \widehat{\Phi}, \widehat{\Xi})$, and the hyperparameters (ϕ, ξ) .
 - 2: Apply the forward–backward algorithm to $r_{\mathcal{H}}(\mathcal{H})$ in Eq. (4.79) and compute $C_{\mathcal{H}}$.
 - 3: Apply (substitute the right-hand side into the left-hand side) Eqs. (4.74), (4.75), (4.76), (4.77), and (4.78) to update $\widehat{\Phi}$, and $\widehat{\Xi}$.
 - 4: Apply Eqs.(4.81) and (4.82) to update ϕ and ξ , respectively.
 - 5: Evaluate the free energy (4.80).
 - 6: Iterate Steps 2 through 5 until convergence (until the energy decrease becomes smaller than a threshold).
-

Similarly, we obtain the following update rule of the hyperparameter ξ :

$$\xi^{\text{new}} = \max \left(0, \xi^{\text{old}} - \frac{KM(\Psi(\xi) - \Psi(M\xi)) - \sum_{k=1}^K \sum_{m=1}^M (\Psi(\widehat{\xi}_{k,m}) - \Psi(\sum_{m'=1}^M \widehat{\xi}_{k,m'})))}{KM(\Psi^{(1)}(\xi) - M\Psi^{(1)}(M\xi))} \right). \quad (4.82)$$

Let

$$\widehat{\mathcal{S}} = \left\{ \left\langle z_k^{(t)} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})} \right\}_{k=1}^K, \left\{ \left\langle z_l^{(t)} z_k^{(t-1)} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})} \right\}_{k,l=1}^K \right\}_{t=1}^T,$$

$\widehat{\Phi} = \{\widehat{\phi}_{k,l}\}_{k,l=1}^K$, and $\widehat{\Xi} = \{\widehat{\xi}_{k,m}\}_{k,m=1}^{K,M}$ be the sets of variational parameters. The EVB learning for the HMM is summarized in Algorithm 10. If the prior hyperparameters are fixed, and Step 4 in the algorithm is omitted, the algorithm reduces to the (nonempirical) VB learning algorithm.

4.2.3 Probabilistic Context-Free Grammars

In this subsection, we discuss *probabilistic context-free grammars (PCFGs)*, which have been used for more complex sequence modeling applications than those with the Markov assumption in natural language processing, bioinformatics, and so on (Durbin et al., 1998). Without loss of generality, we can assume that the grammar is written by the Chomsky normal form. Let the model have K nonterminal symbols and M terminal symbols. The observation sequence of length L is written by $X = (x^{(1)}, \dots, x^{(L)}) \in \{e_1, \dots, e_M\}^L$. Then, the statistical model is defined by

$$p(X|\mathbf{w}) = \sum_{Z \in T(X)} p(X, Z|\mathbf{w}), \quad (4.83)$$

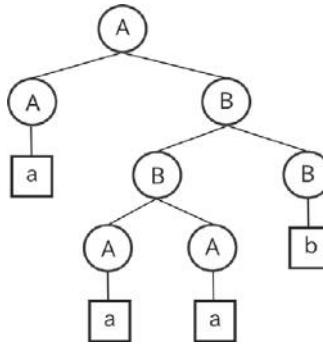


Figure 4.3 Derivation tree of PCFG. A and B are nonterminal symbols and a and b are terminal symbols.

$$\begin{aligned}
 p(X, Z | w) &= \prod_{i,j,k=1}^K (a_{i \rightarrow jk})^{c_{i \rightarrow jk}^Z} \prod_{l=1}^L \prod_{i=1}^K \prod_{m=1}^M (b_{i \rightarrow m})^{\tilde{z}_i^{(l)} x_m^{(l)}}, \\
 w &= \{\{a_i\}_{i=1}^K, \{b_i\}_{i=1}^K\}, \\
 a_i &= \{a_{i \rightarrow jk}\}_{j,k=1}^K \quad (1 \leq i \leq K), \\
 b_i &= \{b_{i \rightarrow m}\}_{m=1}^M \quad (1 \leq i \leq K),
 \end{aligned}$$

where $T(X)$ is the set of derivation sequences that generate X , and Z corresponds to a tree structure representing a derivation sequence. Figure 4.3 illustrates an example of the derivation tree of a PCFG model. The derivation sequence is summarized by $\{c_{i \rightarrow jk}^Z\}_{i,j,k=1}^K$ and $\{\tilde{z}_i^{(l)}\}_{l=1}^L$, where $c_{i \rightarrow jk}^Z$ denotes the count of the transition rule from the nonterminal symbol i to the pair of nonterminal symbols (j, k) appearing in the derivation sequence Z and $\tilde{z}_i^{(l)} = (\tilde{z}_1^{(l)}, \dots, \tilde{z}_K^{(l)})$ is the indicator of the (nonterminal) symbol generating the l th output symbol of X . Moreover the parameter $a_{i \rightarrow jk}$ represents the probability that the nonterminal symbol i emits the pair of nonterminal symbols (j, k) and $b_{i \rightarrow m}$ represents the probability that the nonterminal symbol i emits the terminal symbol m . The parameters, $\{\{a_i\}_{i=1}^K, \{b_i\}_{i=1}^K\}$, have constraints

$$a_{i \rightarrow ii} = 1 - \sum_{(j,k) \neq (i,i)} a_{i \rightarrow jk}, \quad b_{i \rightarrow M} = 1 - \sum_{m=1}^{M-1} b_{i \rightarrow m},$$

i.e., $a_i \in \Delta^{K^2-1}$ and $b_i \in \Delta^{M-1}$, respectively.

Let $\mathcal{D} = \{X^{(1)}, \dots, X^{(N)}\}$ be a given training corpus and $\mathcal{H} = \{Z^{(1)}, \dots, Z^{(N)}\}$ be the corresponding hidden derivation sequences. The log-likelihood for the complete sample $\{\mathcal{D}, \mathcal{H}\}$ is given by

$$\log p(\mathcal{D}, \mathcal{H} | \mathbf{w}) = \sum_{n=1}^N \left[\sum_{i,j,k=1}^K c_{i \rightarrow jk}^{Z^{(n)}} \log a_{i \rightarrow jk} + \sum_{l=1}^L \sum_{i=1}^K \sum_{m=1}^M \bar{z}_i^{(n,l)} x_m^{(n,l)} \log b_{i \rightarrow m} \right],$$

where $x^{(n,l)}$ and $\bar{z}^{(n,l)}$ are the indicators of the l th output symbol and the nonterminal symbol generating the l th output in the n th sequences, $X^{(n)}$ and $Z^{(n)}$, respectively.

We now turn to the VB learning for PCFGs (Kurihara and Sato, 2004). We assume that the prior distributions of parameters $\{\mathbf{a}_i\}_{i=1}^K$ and $\{\mathbf{b}_i\}_{i=1}^K$ are the Dirichlet distributions with hyperparameters ϕ and ξ :

$$p(\{\mathbf{a}_i\}_{i=1}^K | \phi) = \prod_{i=1}^K \text{Dirichlet}_{K^2}(\mathbf{a}_i; (\phi, \dots, \phi)^\top), \quad (4.84)$$

$$p(\{\mathbf{b}_i\}_{i=1}^K | \xi) = \prod_{i=1}^K \text{Dirichlet}_M(\mathbf{b}_i; (\xi, \dots, \xi)^\top). \quad (4.85)$$

VB Posterior for PCFGs

We define the expected sufficient statistics as follows:

$$\bar{N}_{i \rightarrow jk}^z = \sum_{n=1}^N \sum_{l=1}^L \langle c_{i \rightarrow jk}^{Z^{(n)}} \rangle_{r_z(Z^{(n)})}, \quad (4.86)$$

$$\bar{N}_i^z = \sum_{j,k=1}^K \bar{N}_{i \rightarrow jk}^z,$$

$$\bar{N}_{i \rightarrow m}^x = \sum_{n=1}^N \sum_{l=1}^L \langle \bar{z}_i^{(n,l)} \rangle_{r_z(Z^{(n)})} x_m^{(n,l)}, \quad (4.87)$$

$$\bar{N}_i^x = \sum_{m=1}^M \bar{N}_{i \rightarrow m}^x.$$

Then the VB posteriors of the parameters are given by

$$r_w(\mathbf{w}) = r_a(\{\mathbf{a}_i\}_{i=1}^K) r_b(\{\mathbf{b}_i\}_{i=1}^K),$$

$$r_a(\{\mathbf{a}_i\}_{i=1}^K) = \prod_{i=1}^K \text{Dirichlet}_{K^2}(\mathbf{a}_i; (\widehat{\phi}_{i \rightarrow 11}, \dots, \widehat{\phi}_{i \rightarrow KK})^\top), \quad (4.88)$$

$$r_b(\{\mathbf{b}_i\}_{i=1}^K) = \prod_{i=1}^K \text{Dirichlet}_M(\mathbf{b}_i; (\widehat{\xi}_{i \rightarrow 1}, \dots, \widehat{\xi}_{i \rightarrow M})^\top), \quad (4.89)$$

where

$$\widehat{\phi}_{i \rightarrow jk} = \bar{N}_{i \rightarrow jk}^z + \phi, \quad (4.90)$$

$$\widehat{\xi}_{i \rightarrow m} = \bar{N}_{i \rightarrow m}^x + \xi. \quad (4.91)$$

The VB posteriors of the hidden variables are given by

$$\begin{aligned} r_{\mathcal{H}}(\mathcal{H}) &= \prod_{n=1}^N r_z(\mathbf{Z}^{(n)}), \\ r_z(\mathbf{Z}^{(n)}) &= \frac{1}{C_{\mathbf{Z}^{(n)}}} \exp(\gamma_{\mathbf{Z}^{(n)}}), \\ \gamma_{\mathbf{Z}^{(n)}} &= \sum_{i,j,k=1}^K c_{i \rightarrow jk}^{Z^{(n)}} \langle \log a_{i \rightarrow jk} \rangle_{r_a(\{\mathbf{a}_i\}_{i=1}^K)} \\ &\quad + \sum_{l=1}^L \sum_{i=1}^K \sum_{m=1}^M \tilde{z}_i^{(n,l)} x_m^{(n,l)} \langle \log b_{i \rightarrow m} \rangle_{r_b(\{\mathbf{b}_i\}_{i=1}^K)}, \end{aligned} \quad (4.92)$$

where $C_{\mathbf{Z}^{(n)}} = \sum_{\mathbf{Z} \in T(X^{(n)})} \exp(\gamma_{\mathbf{Z}})$ is the normalizing constant and

$$\begin{aligned} \langle \log a_{i \rightarrow jk} \rangle_{r_a(\{\mathbf{a}_i\}_{i=1}^K)} &= \Psi(\widehat{\phi}_{i \rightarrow jk}) - \Psi\left(\sum_{j'=1}^K \sum_{k'=1}^K \widehat{\phi}_{i \rightarrow j'k'}\right), \\ \langle \log b_{i \rightarrow m} \rangle_{r_b(\{\mathbf{b}_i\}_{i=1}^K)} &= \Psi(\widehat{\xi}_{i \rightarrow m}) - \Psi\left(\sum_{m'=1}^M \widehat{\xi}_{i \rightarrow m'}\right). \end{aligned}$$

All the expected sufficient statistics and $C_{\mathbf{Z}^{(n)}}$ can be efficiently computed by the *inside–outside algorithm* (Kurihara and Sato, 2004). The free energy, after the substitution of Eq. (4.3), is given by

$$\begin{aligned} F &= \sum_{i=1}^K \left\{ \log \left(\frac{\Gamma(\sum_{j,k=1}^K \widehat{\phi}_{i \rightarrow jk})}{\prod_{j,k=1}^K \Gamma(\widehat{\phi}_{i \rightarrow jk})} \right) \right. \\ &\quad + \sum_{j,k=1}^K (\widehat{\phi}_{i \rightarrow jk} - \phi) (\Psi(\widehat{\phi}_{i \rightarrow jk}) - \Psi(\sum_{j',k'=1}^K \widehat{\phi}_{i \rightarrow j'k'})) \Big) \\ &\quad \left. + \log \left(\frac{\Gamma(\sum_{m=1}^M \widehat{\xi}_{i \rightarrow m})}{\prod_{m=1}^M \Gamma(\widehat{\xi}_{i \rightarrow m})} \right) + \sum_{m=1}^M (\widehat{\xi}_{i \rightarrow m} - \xi) (\Psi(\widehat{\xi}_{i \rightarrow m}) - \Psi(\sum_{m'=1}^M \widehat{\xi}_{i \rightarrow m'})) \right\} \\ &\quad - K \log \left(\frac{\Gamma(K^2 \phi)}{(\Gamma(\phi))^{K^2}} \right) - K \log \left(\frac{\Gamma(M \xi)}{(\Gamma(\xi))^M} \right) - \sum_{n=1}^N \log C_{\mathbf{Z}^{(n)}}. \end{aligned} \quad (4.93)$$

The following update rules for the EVB learning of the hyperparameters ϕ and ξ are obtained similarly to the HMM in Eqs. (4.81) and (4.82):

$$\phi^{\text{new}} = \max \left(0, \phi^{\text{old}} - \frac{K^3 (\Psi(\phi) - \Psi(K^2 \phi)) - \sum_{i=1}^K \sum_{j,k=1}^K (\Psi(\widehat{\phi}_{i \rightarrow jk}) - \Psi(\sum_{j',k'=1}^K \widehat{\phi}_{i \rightarrow j'k'}))}{K^3 (\Psi^{(1)}(\phi) - K^2 \Psi^{(1)}(K^2 \phi))} \right), \quad (4.94)$$

$$\xi^{\text{new}} = \max \left(0, \xi^{\text{old}} - \frac{KM(\Psi(\xi) - \Psi(M \xi)) - \sum_{i=1}^K \sum_{m=1}^M (\Psi(\widehat{\xi}_{i \rightarrow m}) - \Psi(\sum_{m'=1}^M \widehat{\xi}_{i \rightarrow m'}))}{KM(\Psi^{(1)}(\xi) - M \Psi^{(1)}(M \xi))} \right). \quad (4.95)$$

Algorithm 11 EVB learning for probabilistic context-free grammar.

-
- 1: Initialize the variational parameters $(\widehat{\mathcal{S}}, \widehat{\Phi}, \widehat{\Xi})$ and the hyperparameters (ϕ, ξ) .
 - 2: Apply the inside–outside algorithm to $r_z(\mathbf{Z}^{(n)})$ in Eq. (4.92) and compute $C_{\mathbf{Z}^{(n)}}$ for $n = 1, \dots, N$.
 - 3: Apply (substitute the right-hand side into the left-hand side) Eqs. (4.86), (4.87), (4.90), and (4.91) to update $\widehat{\Phi}$, and $\widehat{\Xi}$.
 - 4: Apply Eqs. (4.94) and (4.95) to update ϕ and ξ , respectively.
 - 5: Evaluate the free energy (4.93).
 - 6: Iterate Steps 2 through 5 until convergence (until the energy decrease becomes smaller than a threshold).
-

Let

$$\widehat{\mathcal{S}} = \left\{ \left\{ \langle c_{i \rightarrow jk}^{(n)} \rangle_{r_z(\mathbf{Z}^{(n)})} \right\}_{i,j,k=1}^K, \left\{ \langle \bar{z}_i^{(n,l)} \rangle_{r_z(\mathbf{Z}^{(n)})} \right\}_{l=1}^L \right\}_{n=1}^N,$$

$\widehat{\Phi} = \{\widehat{\phi}_{i \rightarrow jk}\}_{i,j,k=1}^K$, and $\widehat{\Xi} = \{\widehat{\xi}_{i \rightarrow m}\}_{i,m=1}^{K,M}$ be the sets of variational parameters. The EVB learning for the PCFG model is summarized in Algorithm 11. If the prior hyperparameters are fixed and Step 4 in the algorithm is omitted, the algorithm reduces to the (nonempirical) VB learning algorithm.

4.2.4 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a generative model successfully used in various applications such as text analysis (Blei et al., 2003), image analysis (Li and Perona, 2005), genomics (Bicego et al., 2010; Chen et al., 2010), human activity analysis (Huynh et al., 2008), and collaborative filtering (Krestel et al., 2009; Purushotham et al., 2012). Given word occurrences of documents in a corpora, LDA expresses each document as a mixture of multinomial distributions, each of which is expected to capture a *topic*. The extracted topics provide bases in a low-dimensional feature space, in which each document is compactly represented. This topic expression was shown to be useful for solving various tasks, including classification (Li and Perona, 2005), retrieval (Wei and Croft, 2006), and recommendation (Krestel et al., 2009).

In this subsection, we introduce the VB learning for tractable inference in the LDA model. Suppose that we observe M documents, each of which consists of $N^{(m)}$ words. Each word is included in a vocabulary with size L . We assume that each word is associated with one of the H topics, which is not observed.

We express the word occurrence by an L -dimensional indicator vector \mathbf{w} , where one of the entries is equal to one and the others are equal to zero. Similarly, we express the topic occurrence as an H -dimensional indicator vector \mathbf{z} . We define the following functions that give the item numbers chosen by \mathbf{w} and \mathbf{z} , respectively:

$$\hat{l}(\mathbf{w}) = l \text{ if } w_l = 1 \text{ and } w_{l'} = 0 \text{ for } l' \neq l,$$

$$\hat{h}(\mathbf{z}) = h \text{ if } z_h = 1 \text{ and } z_{h'} = 0 \text{ for } h' \neq h.$$

In the LDA model (Blei et al., 2003), the word occurrence $\mathbf{w}^{(n,m)}$ of the n th position in the m th document is assumed to follow the multinomial distribution:

$$p(\mathbf{w}^{(n,m)}|\boldsymbol{\Theta}, \mathbf{B}) = \prod_{l=1}^L \left((\mathbf{B}\boldsymbol{\Theta}^\top)_{l,m} \right)^{w_l^{(n,m)}} = (\mathbf{B}\boldsymbol{\Theta}^\top)_{\hat{l}(\mathbf{w}^{(n,m)}),m}, \quad (4.96)$$

where $\boldsymbol{\Theta} \in [0, 1]^{M \times H}$ and $\mathbf{B} \in [0, 1]^{L \times H}$ are parameter matrices to be estimated. The rows of $\boldsymbol{\Theta} = (\tilde{\boldsymbol{\theta}}_1, \dots, \tilde{\boldsymbol{\theta}}_M)^\top$ and the columns of $\mathbf{B} = (\beta_1, \dots, \beta_H)$ are probability mass vectors that sum up to one. That is, $\tilde{\boldsymbol{\theta}}_m \in \Delta^{H-1}$ is the topic distribution of the m th document, and $\beta_h \in \Delta^{L-1}$ is the word distribution of the h th topic.

Suppose that we observe the data $\mathcal{D} = \{\{\mathbf{w}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M$. Given the topic occurrence latent variable $\mathbf{z}^{(n,m)}$, the complete likelihood for each word is written as

$$p(\mathbf{w}^{(n,m)}, \mathbf{z}^{(n,m)}|\boldsymbol{\Theta}, \mathbf{B}) = p(\mathbf{w}^{(n,m)}|\mathbf{z}^{(n,m)}, \mathbf{B})p(\mathbf{z}^{(n,m)}|\boldsymbol{\Theta}), \quad (4.97)$$

$$\text{where } p(\mathbf{w}^{(n,m)}|\mathbf{z}^{(n,m)}, \mathbf{B}) = \prod_{l=1}^L \prod_{h=1}^H (B_{l,h})^{w_l^{(n,m)} z_h^{(n,m)}}, \quad p(\mathbf{z}^{(n,m)}|\boldsymbol{\Theta}) = \prod_{h=1}^H (\theta_{m,h})^{z_h^{(n,m)}}.$$

We assume the Dirichlet priors on $\boldsymbol{\Theta}$ and \mathbf{B} :

$$p(\boldsymbol{\Theta}|\boldsymbol{\alpha}) = \prod_{m=1}^M \text{Dirichlet}_H(\tilde{\boldsymbol{\theta}}_m; (\alpha_1, \dots, \alpha_H)^\top), \quad (4.98)$$

$$p(\mathbf{B}|\boldsymbol{\eta}) = \prod_{h=1}^H \text{Dirichlet}_L(\beta_h; (\eta_1, \dots, \eta_L)^\top), \quad (4.99)$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are hyperparameters that control the prior sparsity.

VB Posterior for LDA

For the set of all hidden variables $\mathcal{H} = \{\{\mathbf{z}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M$ and the parameter $\mathbf{w} = (\boldsymbol{\Theta}, \mathbf{B})$, we assume that our approximate posterior is factorized as Eq. (4.1). Thus, the update rule (4.2) yields the further factorization $r_{\Theta, B}(\boldsymbol{\Theta}, \mathbf{B}) = r_\Theta(\boldsymbol{\Theta})r_B(\mathbf{B})$ and the following update rules:

$$r_{\Theta}(\boldsymbol{\Theta}) \propto p(\boldsymbol{\Theta}|\boldsymbol{\alpha}) \langle \log p(\mathcal{D}, \mathcal{H}|\boldsymbol{\Theta}, \mathbf{B}) \rangle_{r_B(\mathbf{B})r_{\mathcal{H}}(\mathcal{H})}, \quad (4.100)$$

$$r_B(\mathbf{B}) \propto p(\mathbf{B}|\boldsymbol{\eta}) \langle \log p(\mathcal{D}, \mathcal{H}|\boldsymbol{\Theta}, \mathbf{B}) \rangle_{r_{\Theta}(\boldsymbol{\Theta})r_{\mathcal{H}}(\mathcal{H})}. \quad (4.101)$$

Define the expected sufficient statistics as

$$\bar{N}_h^{(m)} = \sum_{n=1}^{N^{(m)}} \langle z_h^{(n,m)} \rangle_{r_z(\{\{z^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M)}, \quad (4.102)$$

$$\bar{W}_{l,h} = \sum_{m=1}^M \sum_{n=1}^{N^{(m)}} w_l^{(n,m)} \langle z_h^{(n,m)} \rangle_{r_z(\{\{z^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M)}. \quad (4.103)$$

Then, the VB posterior distribution is given by the Dirichlet distributions:

$$r_{\Theta}(\boldsymbol{\Theta}) = \prod_{m=1}^M \text{Dirichlet}(\bar{\theta}_m; \bar{\alpha}_m), \quad (4.104)$$

$$r_B(\mathbf{B}) = \prod_{h=1}^H \text{Dirichlet}(\boldsymbol{\beta}_h; \bar{\eta}_h), \quad (4.105)$$

where the variational parameters satisfy

$$\bar{\alpha}_{m,h} = (\bar{\alpha}_m)_h = \bar{N}_h^{(m)} + \alpha_h, \quad (4.106)$$

$$\bar{\eta}_{l,h} = (\bar{\eta}_h)_l = \bar{W}_{l,h} + \eta_l. \quad (4.107)$$

From the update rule (4.3), the VB posterior distribution of latent variables is given by the multinomial distribution:

$$r_z(\{\{z^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M) = \prod_{m=1}^M \prod_{n=1}^{N^{(m)}} \text{Multinomial}_{H,1}(z^{(n,m)}; \bar{z}^{(n,m)}), \quad (4.108)$$

where the variational parameter $\bar{z}^{(n,m)} \in \Delta^{H-1}$ is

$$\bar{z}_h^{(n,m)} = \frac{\bar{z}_h^{(n,m)}}{\sum_{h'=1}^H \bar{z}_{h'}^{(n,m)}} \quad (4.109)$$

for

$$\bar{z}_h^{(n,m)} = \exp \left(\langle \log \Theta_{m,h} \rangle_{r_{\Theta}(\boldsymbol{\Theta})} + \sum_{l=1}^L w_l^{(n,m)} \langle \log B_{l,h} \rangle_{r_B(\mathbf{B})} \right). \quad (4.110)$$

We also have

$$\langle z_h^{(n,m)} \rangle_{r_z(\{\{z^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M)} = \bar{z}_h^{(n,m)},$$

$$\langle \log \Theta_{m,h} \rangle_{r_{\Theta}(\boldsymbol{\Theta})} = \Psi(\bar{\alpha}_{m,h}) - \Psi\left(\sum_{h'=1}^H \bar{\alpha}_{m,h'}\right),$$

$$\langle \log B_{l,h} \rangle_{r_B(\mathbf{B})} = \Psi(\bar{\eta}_{l,h}) - \Psi\left(\sum_{l'=1}^L \bar{\eta}_{l',h}\right).$$

Iterating Eqs. (4.106), (4.107), and (4.110) provides a local minimum of the free energy, which is given as a function of the variational parameters by

$$\begin{aligned}
F = & \sum_{m=1}^M \left(\log \left(\frac{\Gamma(\sum_{h=1}^H \widehat{\alpha}_{m,h})}{\prod_{h=1}^H \Gamma(\widehat{\alpha}_{m,h})} \right) - \log \left(\frac{\Gamma(\sum_{h=1}^H \alpha_h)}{\prod_{h=1}^H \Gamma(\alpha_h)} \right) \right) \\
& + \sum_{h=1}^H \left(\log \left(\frac{\Gamma(\sum_{l=1}^L \widehat{\eta}_{l,h})}{\prod_{l=1}^L \Gamma(\widehat{\eta}_{l,h})} \right) - \log \left(\frac{\Gamma(\sum_{l=1}^L \eta_l)}{\prod_{l=1}^L \Gamma(\eta_l)} \right) \right) \\
& + \sum_{m=1}^M \sum_{h=1}^H \left(\widehat{\alpha}_{m,h} - (\bar{N}_h^{(m)} + \alpha_h) \right) \left(\Psi(\widehat{\alpha}_{m,h}) - \Psi(\sum_{h'=1}^H \widehat{\alpha}_{m,h'}) \right) \\
& + \sum_{h=1}^H \sum_{l=1}^L \left(\widehat{\eta}_{l,h} - (\bar{W}_{l,h} + \eta_l) \right) \left(\Psi(\widehat{\eta}_{l,h}) - \Psi(\sum_{l'=1}^L \widehat{\eta}_{l',h}) \right) \\
& + \sum_{m=1}^M \sum_{n=1}^{N^{(m)}} \sum_{h=1}^H \widehat{z}_h^{(n,m)} \log \widehat{z}_h^{(n,m)}. \tag{4.111}
\end{aligned}$$

The partial derivatives of the free energy with respect to $(\boldsymbol{\alpha}, \boldsymbol{\eta})$ are computed as follows:

$$\begin{aligned}
\frac{\partial F}{\partial \alpha_h} = & M \left(\Psi(\alpha_h) - \Psi(\sum_{h'=1}^H \alpha_{h'}) \right) \\
& - \sum_{m=1}^M \left(\Psi(\widehat{\alpha}_{m,h}) - \Psi(\sum_{h'=1}^H \widehat{\alpha}_{m,h'}) \right), \tag{4.112}
\end{aligned}$$

$$\frac{\partial^2 F}{\partial \alpha_h \partial \alpha_{h'}} = M \left(\delta_{h,h'} \Psi^{(1)}(\alpha_h) - \Psi^{(1)}(\sum_{h''=1}^H \alpha_{h''}) \right), \tag{4.113}$$

$$\begin{aligned}
\frac{\partial F}{\partial \eta_l} = & H \left(\Psi(\eta_l) - \Psi(\sum_{l'=1}^L \eta_{l'}) \right) \\
& - \sum_{h=1}^H \left(\Psi(\widehat{\eta}_{l,h}) - \Psi(\sum_{l'=1}^L \widehat{\eta}_{l',h}) \right), \tag{4.114}
\end{aligned}$$

$$\frac{\partial^2 F}{\partial \eta_l \partial \eta_{l'}} = H \left(\delta_{l,l'} \Psi^{(1)}(\eta_l) - \Psi^{(1)}(\sum_{l''=1}^L \eta_{l''}) \right), \tag{4.115}$$

where $\delta_{n,n'}$ is the *Kronecker delta*. Thus, we have the following Newton–Raphson steps to update the hyperparameters:

$$\boldsymbol{\alpha}^{\text{new}} = \mathbf{max} \left(\mathbf{0}, \boldsymbol{\alpha}^{\text{old}} - \left(\frac{\partial^2 F}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top} \right)^{-1} \frac{\partial F}{\partial \boldsymbol{\alpha}} \right), \tag{4.116}$$

$$\boldsymbol{\eta}^{\text{new}} = \mathbf{max} \left(\mathbf{0}, \boldsymbol{\eta}^{\text{old}} - \left(\frac{\partial^2 F}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top} \right)^{-1} \frac{\partial F}{\partial \boldsymbol{\eta}} \right), \tag{4.117}$$

where $\mathbf{max}(\cdot)$ is the max operator applied elementwise.

Algorithm 12 EVB learning for latent Dirichlet allocation.

- 1: Initialize the variational parameters $(\{\widehat{\mathbf{z}}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M, \{\widehat{\boldsymbol{\alpha}}_m\}_{m=1}^M, \{\widehat{\boldsymbol{\eta}}_h\}_{h=1}^H)$, and the hyperparameters $(\boldsymbol{\alpha}, \boldsymbol{\eta})$.
 - 2: Apply (substitute the right-hand side into the left-hand side) Eqs. (4.110), (4.109), (4.102), (4.103), (4.106), and (4.107) to update $\{\widehat{\mathbf{z}}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M, \{\widehat{\boldsymbol{\alpha}}_m\}_{m=1}^M$, and $\{\widehat{\boldsymbol{\eta}}_h\}_{h=1}^H$.
 - 3: Apply Eqs. (4.116) and (4.117) to update $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, respectively.
 - 4: Evaluate the free energy (4.111).
 - 5: Iterate Steps 2 through 4 until convergence (until the energy decrease becomes smaller than a threshold).
-

The EVB learning for LDA is summarized in Algorithm 12. If the prior hyperparameters are fixed and Step 3 in the algorithm is omitted, the algorithm reduces to the (nonempirical) VB learning algorithm.

We can also apply partially Bayesian (PB) learning by approximating the posterior of $\boldsymbol{\Theta}$ or \mathbf{B} by the delta function (see Section 2.2.2). We call it PB-A learning if $\boldsymbol{\Theta}$ is marginalized and \mathbf{B} is point-estimated, and PB-B learning if \mathbf{B} is marginalized and $\boldsymbol{\Theta}$ is point-estimated. Note that the original VB algorithm for LDA proposed by Blei et al. (2003) corresponds to PB-A learning in our terminology. MAP learning, where both of $\boldsymbol{\Theta}$ and \mathbf{B} are point-estimated, corresponds to the *probabilistic latent semantic analysis (pLSA)* (Hofmann, 2001), if we assume the flat prior $\alpha_h = \eta_l = 1$ (Girolami and Kaban, 2003).

5

VB Algorithm under No Conjugacy

As discussed in Section 2.1.7, there are practical combinations of a model and a prior where conjugacy is no longer available. In this chapter, as a method for addressing nonconjugacy, we demonstrate *local variational approximation (LVA)*, also known as *direct site bounding*, for logistic regression and a sparsity-inducing prior (Jaakkola and Jordan, 2000; Girolami, 2001; Bishop, 2006; Seeger, 2008, 2009). Then we describe a general framework of LVA based on convex functions by using the associated Bregman divergence (Watanabe et al., 2011).

5.1 Logistic Regression

Let $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$ be the N observations of the binary response variable $y^{(n)} \in \{0, 1\}$ and the input vector $\mathbf{x}^{(n)} \in \mathbb{R}^M$. The *logistic regression* model assumes the following Bernoulli model over $\mathbf{y} = (y^{(1)}, y^{(2)}, \dots, y^{(N)})^\top$ given $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\}$:

$$p(\mathbf{y}|X, \mathbf{w}) = \prod_{n=1}^N \exp\left(y^{(n)}(\mathbf{w}^\top \mathbf{x}^{(n)}) - \log(1 + e^{\mathbf{w}^\top \mathbf{x}^{(n)}})\right). \quad (5.1)$$

Let us consider the Bayesian learning of the parameter \mathbf{w} assuming the Gaussian prior distribution:

$$p(\mathbf{w}) = \text{Gauss}_M(\mathbf{w}; \mathbf{w}_0, \mathbf{S}_0^{-1}),$$

where \mathbf{S}_0 and \mathbf{w}_0 are the hyperparameters.

Gaussian approximations for the posterior distribution $p(\mathbf{w}|\mathcal{D}) \propto p(\mathbf{w}, \mathbf{y}|X)$ are obtained by LVA based on the facts that $-\log(e^{\sqrt{h}/2} + e^{-\sqrt{h}/2})$ is a convex function of h and that $\log(1 + e^g)$ is a convex function of g . More specifically,

because $\phi(h(w)) = -\log(e^{\sqrt{w^2}/2} + e^{-\sqrt{w^2}/2})$ is a convex function of $h(w) = w^2$ and $\psi(g(w)) = \log(1 + e^w)$ is a convex function of $g(w) = w$, they are bounded from below by their tangents at $h(\xi) = \xi^2$ and $g(\eta) = \eta$, respectively:

$$\begin{aligned} -\log\left(e^{\sqrt{w^2}/2} + e^{-\sqrt{w^2}/2}\right) &\geq -\log\left(e^{\sqrt{\xi^2}/2} + e^{-\sqrt{\xi^2}/2}\right) - (w^2 - \xi^2)\frac{\tanh(\xi/2)}{4\xi}, \\ \log(1 + e^w) &\geq \log(1 + e^\eta) + (w - \eta)\frac{e^\eta}{1 + e^\eta}. \end{aligned}$$

By substituting these bounds into the likelihood (5.1), we obtain the following bounds on $p(\mathbf{w}, \mathbf{y}|X)$:

$$\underline{p}(\mathbf{w}; \boldsymbol{\xi}) \leq p(\mathbf{w}, \mathbf{y}|X) \leq \bar{p}(\mathbf{w}; \boldsymbol{\eta}),$$

where

$$\begin{aligned} \underline{p}(\mathbf{w}; \boldsymbol{\xi}) &\equiv p(\mathbf{w}) \prod_{n=1}^N \exp\left(\left(y^{(n)} - \frac{1}{2}\right) \mathbf{w}^\top \mathbf{x}^{(n)} \right. \\ &\quad \left. - \theta_n \left\{ (\mathbf{w}^\top \mathbf{x}^{(n)})^2 - h_n \right\} - \log\left(e^{\frac{\sqrt{h_n}}{2}} + e^{-\frac{\sqrt{h_n}}{2}}\right)\right), \\ \bar{p}(\mathbf{w}; \boldsymbol{\eta}) &\equiv p(\mathbf{w}) \prod_{n=1}^N \exp\left((y^{(n)} - \kappa_n) \mathbf{w}^\top \mathbf{x}^{(n)} - b(\kappa_n)\right). \end{aligned}$$

Here we have put

$$\theta_n = \frac{\tanh(\sqrt{h_n}/2)}{4\sqrt{h_n}}, \quad (5.2)$$

$$\kappa_n = \frac{e^{g_n}}{1 + e^{g_n}}, \quad (5.3)$$

and $\{h_n\}_{n=1}^N$ and $\{g_n\}_{n=1}^N$ are the sets of variational parameters defined from $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_M)^\top$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_M)^\top$ by the transformations $h_n = (\boldsymbol{\xi}^\top \mathbf{x}^{(n)})^2$ and $g_n = \boldsymbol{\eta}^\top \mathbf{x}^{(n)}$, respectively. We also defined the binary entropy function by $b(\kappa) = -\kappa \log \kappa - (1 - \kappa) \log(1 - \kappa)$ for $\kappa \in [0, 1]$.

Normalizing these bounds with respect to \mathbf{w} , we approximate the posterior by the Gaussian distributions as

$$\begin{aligned} q_\xi(\mathbf{w}; \boldsymbol{\xi}) &= \text{Gauss}_M(\underline{\mathbf{m}}, \underline{\mathbf{S}}^{-1}), \\ q_\eta(\mathbf{w}; \boldsymbol{\eta}) &= \text{Gauss}_M(\bar{\mathbf{m}}, \bar{\mathbf{S}}^{-1}), \end{aligned}$$

whose mean and precision (inverse-covariance) matrix are respectively given by

$$\begin{aligned} \underline{\mathbf{m}} &= \underline{\mathbf{S}}^{-1} \left\{ S_0 \mathbf{w}_0 + \sum_{n=1}^N (y^{(n)} - 1/2) \mathbf{x}^{(n)} \right\}, \\ \underline{\mathbf{S}} &= S_0 + 2 \sum_{n=1}^N \theta_n \mathbf{x}^{(n)} \mathbf{x}^{(n)\top}, \end{aligned} \quad (5.4)$$

and

$$\begin{aligned}\underline{\mathbf{m}} &= \mathbf{w}_0 + \mathbf{S}_0^{-1} \sum_{n=1}^N (\mathbf{y}^{(n)} - \boldsymbol{\kappa}_n) \mathbf{x}^{(n)}, \\ \underline{\mathbf{S}} &= \mathbf{S}_0.\end{aligned}\quad (5.5)$$

We also obtain the bounds for the marginal likelihood, $\underline{Z}(\boldsymbol{\xi}) \equiv \int \underline{p}(\mathbf{w}; \boldsymbol{\xi}) d\mathbf{w}$ and $\bar{Z}(\boldsymbol{\eta}) \equiv \int \bar{p}(\mathbf{w}; \boldsymbol{\eta}) d\mathbf{w}$. These are respectively given in the forms of free energy bounds as follows:

$$\begin{aligned}\bar{F}(\boldsymbol{\xi}) &\equiv -\log \underline{Z}(\boldsymbol{\xi}) \\ &= \frac{1}{2} \log |\underline{\mathbf{S}}| - \frac{1}{2} \log |\mathbf{S}_0| + \frac{\mathbf{w}_0^\top \mathbf{S}_0 \mathbf{w}_0}{2} - \frac{\underline{\mathbf{m}}^\top (\underline{\mathbf{S}}) \underline{\mathbf{m}}}{2} \\ &\quad - \sum_{n=1}^N \left\{ h_n \theta_n - \log \left(2 \cosh \left(\frac{\sqrt{h_n}}{2} \right) \right) \right\},\end{aligned}\quad (5.6)$$

and

$$\begin{aligned}\underline{F}(\boldsymbol{\eta}) &\equiv -\log \bar{Z}(\boldsymbol{\eta}) \\ &= \frac{\mathbf{w}_0^\top \mathbf{S}_0 \mathbf{w}_0}{2} - \frac{\bar{\mathbf{m}}^\top \mathbf{S}_0 \bar{\mathbf{m}}}{2} + \sum_{n=1}^N b(\kappa_n).\end{aligned}$$

We optimize the free energy bounds to determine the variational parameters. As will be discussed generally in Section 5.3.2, to decrease the upper-bound $\bar{F}(\boldsymbol{\xi})$, the EM algorithm is available, which instead maximizes

$$\langle \log \underline{p}(\mathbf{w}; \boldsymbol{\xi}) \rangle_{q_\xi(\mathbf{w}; \boldsymbol{\xi}_0)},$$

where the expectation is taken with respect to the approximate posterior before updating with the variational parameters given by $\boldsymbol{\xi}_0$. The update rule of the variational parameters is specifically given by

$$\begin{aligned}h_n &= \langle (\mathbf{w}^\top \mathbf{x}^{(n)})^2 \rangle_{q_\xi(\mathbf{w}; \boldsymbol{\xi}_0)} \\ &= \mathbf{x}^{(n)\top} (\underline{\mathbf{S}}^{-1} + \underline{\mathbf{m}} \underline{\mathbf{m}}^\top) \mathbf{x}^{(n)},\end{aligned}\quad (5.7)$$

where $\underline{\mathbf{m}}$ and $\underline{\mathbf{S}}^{-1}$ are the mean and covariance matrix of $q_\xi(\mathbf{w}; \boldsymbol{\xi}_0)$.

We can use the following gradient for the maximization of the lower-bound $\underline{F}(\boldsymbol{\eta})$:

$$\begin{aligned}\frac{\partial \underline{F}(\boldsymbol{\eta})}{\partial \kappa_n} &= \langle \mathbf{w}^\top \mathbf{x}^{(n)} \rangle_{q_\eta(\mathbf{w}; \boldsymbol{\eta})} - \boldsymbol{\eta}^\top \mathbf{x}^{(n)} \\ &= \bar{\mathbf{m}}^\top \mathbf{x}^{(n)} - g_n.\end{aligned}\quad (5.8)$$

The Newton–Raphson step to update $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_N)^\top$ is given by

$$\boldsymbol{\kappa}^{\text{new}} = \boldsymbol{\kappa}^{\text{old}} - \left(\frac{\partial^2 \underline{F}}{\partial \boldsymbol{\kappa} \partial \boldsymbol{\kappa}^\top} \right)^{-1} \frac{\partial \underline{F}}{\partial \boldsymbol{\kappa}},\quad (5.9)$$

Algorithm 13 LVA algorithm for logistic regression.

-
- 1: Initialize the variational parameters $\{h_n\}_{n=1}^N$ and transform them to $\{\theta_n\}_{n=1}^N$ by Eq. (5.2).
 - 2: Compute the approximate posterior mean and covariance matrix $(\underline{\mathbf{m}}, \underline{\mathbf{S}}^{-1})$ by Eq. (5.4).
 - 3: Apply Eq. (5.7) to update $\{h_n\}_{n=1}^N$ and transform them to $\{\theta_n\}_{n=1}^N$ by Eq. (5.2).
 - 4: Evaluate the free energy bound (5.6).
 - 5: Iterate Steps 2 through 4 until convergence (until the decrease of the bound becomes smaller than a threshold).
-

where the (n, n') th entry of the Hessian matrix is given as follows:

$$\frac{\partial^2 \underline{F}(\boldsymbol{\eta})}{\partial \kappa_n \partial \kappa_{n'}} = -\mathbf{x}^{(n)\top} \mathbf{S}_0^{-1} \mathbf{x}^{(n')} - \delta_{n,n'} \left(\frac{1}{\kappa_n} + \frac{1}{1-\kappa_n} \right).$$

The learning algorithm for logistic regression with LVA is summarized in Algorithm 13 in the case of $\underline{F}(\boldsymbol{\xi})$ minimization. To obtain the algorithm for $\underline{F}(\boldsymbol{\eta})$ maximization, the updated variables are replaced with $\{g_n\}_{n=1}^N$, $\{\kappa_n\}_{n=1}^N$, and $(\bar{\mathbf{m}}, \bar{\mathbf{S}}^{-1})$, and the update rule (5.7) in Step 3 is replaced with the Newton–Raphson step (5.9).

Recall the arguments in Section 2.1.7 that the VB posterior $r(\mathbf{w}; \hat{\lambda}) = q(\mathbf{w}; \boldsymbol{\xi})$ in Eq. (2.32) minimizes the upper-bound of the free energy (2.25) and the approximate posterior $r(\mathbf{w}; \hat{\nu}) = q(\mathbf{w}; \boldsymbol{\eta})$ in Eq. (2.58) maximizes the lower-bound of the objective function of EP (2.50). This means that the variational parameters are given by $\hat{\lambda} = (\underline{\mathbf{m}}, \underline{\mathbf{S}}^{-1})$ and $\hat{\nu} = (\bar{\mathbf{m}}, \bar{\mathbf{S}}^{-1})$ in the LVAs for VB and EP, respectively.

5.2 Sparsity-Inducing Prior

Another representative example where a nonconjugate prior is used is the linear regression model with a *sparsity-inducing prior* distribution (Girolami, 2001; Seeger, 2008, 2009). We discuss the linear regression model for i.i.d. data $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$, where for each observation, $\mathbf{x}^{(n)} \in \mathbb{R}^M$ is the input vector and $y^{(n)} \in \mathbb{R}$ is the response. Denoting $\mathbf{y} = (y^{(1)}, \dots, y^{(N)})^\top$ and $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})^\top$, we assume the model,

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \text{Gauss}_N(\mathbf{y}; \mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N),$$

and the following sparsity-inducing prior with the L_β -norm:

$$p(\mathbf{w}) = \prod_{m=1}^M \frac{1}{C_{\beta,\gamma}} \exp(-\gamma|w_m|^\beta), \quad (5.10)$$

where $\gamma > 0$ and $0 < \beta \leq 2$ are the hyperparameters, and $C_{\beta,\gamma} = \frac{2}{\beta} \gamma^{1-1/\beta} \Gamma(1/\beta)$ is the normalizing constant. For $0 < \beta < 2$, the prior has a heavier tail than the Gaussian distribution ($\beta = 2$) and induces sparsity of the coefficients \mathbf{w} .

We apply the following inequality for $w, \xi \in \mathbb{R}$:

$$(w^2)^{\beta/2} \leq \frac{\beta}{2} (\xi^2)^{\frac{\beta}{2}-1} (w^2 - \xi^2),$$

which is obtained from the concavity of the function $f(x) = x^{\beta/2}$ for $x > 0$ and $0 < \beta < 2$. Introducing the variational parameter to each dimension, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_M)^\top$ and bounding the nonconjugate prior (5.10) by this inequality, we have

$$\begin{aligned} p(\mathbf{y}, \mathbf{w} | \mathbf{X}) &= p(\mathbf{y} | \mathbf{X}, \mathbf{w}) p(\mathbf{w}) \\ &\geq \frac{1}{(2\pi\sigma^2)^{N/2} C_{\beta,\gamma}^M} \\ &\quad \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N (\mathbf{y}^{(n)} - \mathbf{w}^\top \mathbf{x}^{(n)})^2 - \frac{\beta\gamma}{2} \sum_{m=1}^M (\xi_m^2)^{\frac{\beta}{2}-1} (w_m^2 - \xi_m^2)\right) \\ &\equiv \underline{p}(\mathbf{w}; \boldsymbol{\xi}). \end{aligned}$$

Normalizing the lower-bound $\underline{p}(\mathbf{w}; \boldsymbol{\xi})$, we obtain a Gaussian approximation to the posterior. This is in effect equivalent to assuming the Gaussian prior for \mathbf{w} :

$$\text{Gauss}_M(\mathbf{w}; \mathbf{0}, \mathbf{S}_\xi^{-1}),$$

where $\mathbf{S}_\xi = \gamma\beta \text{Diag}(\boldsymbol{\xi}^{\beta-1/2})$ for $\boldsymbol{\xi}^{\beta-1/2} \equiv \left((\xi_1^2)^{\frac{\beta}{2}-1}, \dots, (\xi_M^2)^{\frac{\beta}{2}-1}\right)^\top$.

The resulting Gaussian approximation to the posterior is

$$q_\xi(\mathbf{w}; \boldsymbol{\xi}) = \text{Gauss}_M(\mathbf{w}; \underline{\mathbf{m}}, \underline{\mathbf{S}}^{-1}),$$

where

$$\underline{\mathbf{S}} = \mathbf{S}_\xi + \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X}, \quad (5.11)$$

$$\underline{\mathbf{m}} = \frac{1}{\sigma^2} \underline{\mathbf{S}}^{-1} \mathbf{X}^\top \mathbf{y}. \quad (5.12)$$

Algorithm 14 LVA algorithm for sparse linear regression.

-
- 1: Initialize the variational parameters $\{\xi_m^2\}_{m=1}^M$.
 - 2: Compute the approximate posterior mean and covariance matrix $(\underline{m}, \underline{S}^{-1})$ by Eqs. (5.11) and (5.12).
 - 3: Apply Eq. (5.14) to update $\{\xi_m^2\}_{m=1}^M$.
 - 4: Evaluate the free energy bound (5.13).
 - 5: Iterate Steps 2 through 4 until convergence (until the decrease of the bound becomes smaller than a threshold).
-

We also obtain the upper bound for the free energy:

$$\begin{aligned}\bar{F}(\xi) &= -\log \int \underline{p}(\mathbf{w}; \xi) d\mathbf{w} \\ &= \frac{N-M}{2} \log(2\pi) + \frac{\log |\underline{S}|}{2} + M \log C_{\beta, \gamma} - \sum_{n=1}^N \frac{(y^{(n)})^2}{2\sigma^2} + \frac{\gamma\beta}{2} \sum_{m=1}^M (\xi_m^2)^{\beta/2},\end{aligned}\quad (5.13)$$

which is optimized with respect to the variational parameter. The general framework in Section 5.3.2, which corresponds to the EM algorithm, provides the following update rule:

$$\begin{aligned}\xi_m^2 &= \left\langle w_m^2 \right\rangle_{q_\xi(\mathbf{w}; \xi_0)} \\ &= (\underline{S}^{-1})_{mm} + \underline{m}_m^2,\end{aligned}\quad (5.14)$$

where \underline{m} and \underline{S}^{-1} are the mean and covariance matrix of $q_\xi(\mathbf{w}; \xi_0)$.

The learning algorithm for sparse linear regression with this approximation is summarized in Algorithm 14.

This approximation has been applied to the Laplace prior ($\beta = 1$) in Seeger (2008, 2009). LVA for another heavy-tailed distribution, $p(w) \propto \cosh^{-1/\beta}(\beta w)$, is discussed in Girolami (2001), which also bridges the Gaussian distribution ($\beta \rightarrow 0$) and the Laplace distribution ($\beta \rightarrow \infty$).

5.3 Unified Approach by Local VB Bounds

As discussed in Section 2.1.7, LVA for VB and LVA for EP form lower- and upper-bounds of the joint distribution $p(\mathbf{w}, \mathcal{D})$, denoted by $\underline{p}(\mathbf{w}; \xi)$ and $\bar{p}(\mathbf{w}; \eta)$, respectively. If the bounds satisfying

$$\underline{p}(\mathbf{w}; \boldsymbol{\xi}) \leq p(\mathbf{w}, \mathcal{D}), \quad (5.15)$$

$$\bar{p}(\mathbf{w}; \boldsymbol{\eta}) \geq p(\mathbf{w}, \mathcal{D}), \quad (5.16)$$

for all \mathbf{w} and \mathcal{D} are analytically integrable, then by normalizing the bounds instead of $p(\mathbf{w}, \mathcal{D})$, LVAs approximate the posterior distribution by

$$q_{\xi}(\mathbf{w}; \boldsymbol{\xi}) = \frac{\underline{p}(\mathbf{w}; \boldsymbol{\xi})}{\underline{Z}(\boldsymbol{\xi})}, \quad (5.17)$$

$$q_{\eta}(\mathbf{w}; \boldsymbol{\eta}) = \frac{\bar{p}(\mathbf{w}; \boldsymbol{\eta})}{\bar{Z}(\boldsymbol{\eta})}, \quad (5.18)$$

respectively, where $\underline{Z}(\boldsymbol{\xi})$ and $\bar{Z}(\boldsymbol{\eta})$ are the normalization constants defined by

$$\begin{aligned} \underline{Z}(\boldsymbol{\xi}) &= \int \underline{p}(\mathbf{w}; \boldsymbol{\xi}) d\mathbf{w}, \\ \bar{Z}(\boldsymbol{\eta}) &= \int \bar{p}(\mathbf{w}; \boldsymbol{\eta}) d\mathbf{w}. \end{aligned}$$

Here $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are called the variational parameters.

The respective approximations are optimized by estimating the variational parameters, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ so that $\underline{Z}(\boldsymbol{\xi})$ is maximized and $\bar{Z}(\boldsymbol{\eta})$ is minimized since the inequalities

$$\underline{Z}(\boldsymbol{\xi}) \leq Z \leq \bar{Z}(\boldsymbol{\eta}) \quad (5.19)$$

hold by definition, where $Z = p(\mathcal{D})$ is the marginal likelihood.

To consider the respective LVAs in terms of information divergences in later sections, let us introduce the Bayes free energy,

$$F^{\text{Bayes}} \equiv -\log Z,$$

and its lower- and upper-bounds, $\underline{F}(\boldsymbol{\eta}) = -\log \bar{Z}(\boldsymbol{\eta})$ and $\bar{F}(\boldsymbol{\xi}) = -\log \underline{Z}(\boldsymbol{\xi})$. By taking the negative logarithms on both sides of Eq. (5.19), we have

$$\underline{F}(\boldsymbol{\eta}) \leq F^{\text{Bayes}} \leq \bar{F}(\boldsymbol{\xi}). \quad (5.20)$$

Hereafter, we follow the measure of the free energy and adopt the following terminology to refer to respective LVAs (5.18) and (5.17): the lower-bound maximization ($\underline{F}(\boldsymbol{\eta})$ maximization) and the upper-bound minimization ($\bar{F}(\boldsymbol{\xi})$ minimization).

5.3.1 Divergence Measures in LVA

Most of the existing LVA techniques are based on the convexity of the log-likelihood function or the log-prior (Jaakkola and Jordan, 2000; Bishop,

2006; Seeger, 2008, 2009). We describe these cases by using general convex functions, ϕ and ψ , and show that the objective functions,

$$\begin{aligned}\bar{F}(\xi) - F^{\text{Bayes}} &= \log \frac{Z}{\underline{Z}(\xi)} \geq 0, \\ F^{\text{Bayes}} - \underline{F}(\eta) &= \log \frac{\bar{Z}(\eta)}{Z} \geq 0,\end{aligned}$$

to be minimized in the approximations (5.17) and (5.18), are decomposable into the sum of the KL divergence and the expected Bregman divergence.

Let ϕ and ψ be twice differentiable real-valued strictly convex functions and denote by d_ϕ the Bregman divergence associated with the function ϕ (Banerjee et al., 2005):

$$d_\phi(v_1, v_2) = \phi(v_1) - \phi(v_2) - (v_1 - v_2)^\top \nabla \phi(v_2) \geq 0, \quad (5.21)$$

where $\nabla \phi(v_2)$ denotes the gradient vector of ϕ at v_2 .

Let us consider the case where ϕ and ψ are respectively used to form the following bounds of the joint distribution $p(w, \mathcal{D})$:

$$\underline{p}(w; \xi) = p(w, \mathcal{D}) \exp\{-d_\phi(h(w), h(\xi))\}, \quad (5.22)$$

$$\bar{p}(w; \eta) = p(w, \mathcal{D}) \exp\{d_\psi(g(w), g(\eta))\}, \quad (5.23)$$

where h and g are vector-valued functions of w .¹

Eq. (5.22) is interpreted as follows. $\log p(w, \mathcal{D})$ includes a term that prevents analytic integration of $p(w, \mathcal{D})$ with respect to w . If such a term is expressed by the convex function ϕ of some function h transforming w , it is replaced by the tangent hyperplane, $\phi(h(\xi)) + (h(w) - h(\xi))^\top \nabla \phi(h(\xi))$, so that $\log \underline{p}(w; \xi)$ makes a simpler function of w , such as a quadratic function. Remember that if $\log p(w; \xi)$ is quadratic with respect to w , $\underline{p}(w; \xi)$ is analytically integrable by the Gaussian integral.

Rephrased in terms of the convex duality theory (Jordan et al., 1999; Bishop, 2006), $\phi(h(w))$ is replaced by its lower-bound,

$$\phi(h(w)) \geq \phi(h(\xi)) + (h(w) - h(\xi))^\top \nabla \phi(h(\xi)) \quad (5.24)$$

$$= \theta(\xi)^\top h(w) - \tilde{\phi}(\theta(\xi)), \quad (5.25)$$

where we have put $\theta(\xi) = \nabla \phi(h(\xi))$ and

$$\tilde{\phi}(\theta(\xi)) = \theta(\xi)^\top h(\xi) - \phi(h(\xi))$$

$$= \max_h \{\theta(\xi)^\top h - \phi(h)\}$$

¹ The functions g and h (also ψ and ϕ) can be dependent on \mathcal{D} in this discussion. However, we denote them as if they were independent of \mathcal{D} for simplicity. They are actually independent of \mathcal{D} in the examples in Sections 5.1 and 5.2 and in most applications (Bishop, 2006; Seeger, 2008, 2009).

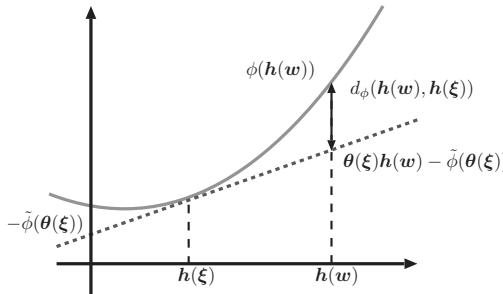


Figure 5.1 Convex function ϕ (solid curve), its tangent (dashed line), and the Bregman divergence (arrow).

is the conjugate function of ϕ . The inequality (5.24) indicates the fact that the convex function ϕ is bounded globally by its tangent at $\mathbf{h}(\xi)$, which is equivalent to the nonnegativity of the Bregman divergence. In Eq. (5.25), the tangent is reparameterized by $\theta(\xi)$, its gradient, instead of the contact point $\mathbf{h}(\xi)$, and its offset is given by $-\tilde{\phi}(\theta(\xi))$. Figure 5.1 illustrates the relationship among the convex function ϕ , its lower-bound, and the Bregman divergence.

We now describe the free energy bounds $\bar{F}(\xi)$ and $\underline{F}(\eta)$ in terms of information divergences. It follows from the definition (5.17) of the approximate posterior distribution that

$$\begin{aligned} \text{KL}(q_\xi(\mathbf{w}; \xi) \| p(\mathbf{w} | \mathcal{D})) &= \int q_\xi(\mathbf{w}; \xi) \log \frac{Z_p(\mathbf{w}; \xi)}{Z(\xi)p(\mathbf{w}, \mathcal{D})} d\mathbf{w} \\ &= \log \frac{Z}{Z(\xi)} - \left\langle d_\phi(\mathbf{h}(\mathbf{w}), \mathbf{h}(\xi)) \right\rangle_{q_\xi(\mathbf{w}; \xi)}. \end{aligned}$$

We have a similar decomposition for $\text{KL}(p(\mathbf{w} | \mathcal{D}) \| q_\eta(\mathbf{w}; \eta))$ as well. Finally, we obtain the following expressions:²

$$\bar{F}(\xi) - F^{\text{Bayes}} = \left\langle d_\phi(\mathbf{h}(\mathbf{w}), \mathbf{h}(\xi)) \right\rangle_{q_\xi} + \text{KL}(q_\xi \| p), \quad (5.26)$$

$$F^{\text{Bayes}} - \underline{F}(\eta) = \left\langle d_\psi(\mathbf{g}(\mathbf{w}), \mathbf{g}(\eta)) \right\rangle_p + \text{KL}(p \| q_\eta). \quad (5.27)$$

Recall that $F^{\text{Bayes}} + \text{KL}(q_\xi \| p) = F$ is the free energy, which is further bounded by $\bar{F}(\lambda, \xi)$ in Eq. (2.25). The expression (5.26) shows that the gap between $\min_{\lambda} \bar{F}(\lambda, \xi)$ and F is the expected Bregman divergence $\left\langle d_\phi(\mathbf{h}(\mathbf{w}), \mathbf{h}(\xi)) \right\rangle_{q_\xi}$.

Recall also that $F^{\text{Bayes}} - \text{KL}(p \| q_\eta) = E$ is the objective function of the EP problem (2.46) and that $\underline{F}(\eta) = -\log \bar{Z}(\eta)$ is obtained as the maximum of its lower-bound, $\max_{\lambda} \underline{E}(\lambda, \eta)$ in Eq. (2.59), under a monotonic transformation.

² Hereafter in this section, we omit the notation “ $(\mathbf{w} | \mathcal{D})$ ” if no confusion is likely.

The expression (5.27) shows that the gap between E and $\max_{\widehat{\nu}} \underline{E}(\widehat{\nu}, \boldsymbol{\eta})$ is expressed by the expected Bregman divergence $\langle d_\psi(\mathbf{g}(\mathbf{w}), \mathbf{g}(\boldsymbol{\eta})) \rangle_p$ while the expectation is taken with respect to the true posterior.

Similarly, we also have the following decompositions:

$$\overline{F}(\boldsymbol{\xi}) - F^{\text{Bayes}} = \langle d_\phi(\mathbf{h}(\mathbf{w}), \mathbf{h}(\boldsymbol{\xi})) \rangle_p - \text{KL}(p \| q_\xi), \quad (5.28)$$

and

$$F^{\text{Bayes}} - \underline{F}(\boldsymbol{\eta}) = \langle d_\psi(\mathbf{g}(\mathbf{w}), \mathbf{g}(\boldsymbol{\eta})) \rangle_{q_\eta} - \text{KL}(q_\eta \| p).$$

Unlike Eqs. (5.26) and (5.27), the KL divergence is subtracted in these expressions. This again implies the affinities of LVAs by \overline{F} minimization and \underline{F} maximization to VB and EP, respectively.

5.3.2 Optimization of Approximations

In this subsection, we show that the conditions for the optimal variational parameters are generally given by the moment matching with respect to $\mathbf{h}(\boldsymbol{\xi})$ and $\mathbf{g}(\boldsymbol{\eta})$.

Optimal Variational Parameters

From Eqs. (5.22) and (5.23), we can see that the approximate posteriors,

$$\begin{aligned} q_\xi(\mathbf{w}; \boldsymbol{\xi}) &\propto p(\mathbf{w}, \mathcal{D}) \exp\{\mathbf{h}(\mathbf{w})^\top \nabla \phi(\mathbf{h}(\boldsymbol{\xi})) - \phi(\mathbf{h}(\mathbf{w}))\} \\ q_\eta(\mathbf{w}; \boldsymbol{\eta}) &\propto p(\mathbf{w}, \mathcal{D}) \exp\{-\mathbf{g}(\mathbf{w})^\top \nabla \psi(\mathbf{g}(\boldsymbol{\eta})) + \psi(\mathbf{g}(\mathbf{w}))\}, \end{aligned} \quad (5.29)$$

are members of the exponential family with natural parameters $\nabla \phi(\mathbf{h}(\boldsymbol{\xi}))$ and $\nabla \psi(\mathbf{g}(\boldsymbol{\eta}))$ (Section 1.2.3). Let

$$\boldsymbol{\theta}(\boldsymbol{\xi}) = \nabla \phi(\mathbf{h}(\boldsymbol{\xi})) \quad \text{and} \quad \boldsymbol{\kappa}(\boldsymbol{\eta}) = \nabla \psi(\mathbf{g}(\boldsymbol{\eta})).$$

The variational parameters are optimized so that $\overline{F}(\boldsymbol{\xi})$ is minimized and $\underline{F}(\boldsymbol{\eta})$ is maximized, respectively. In practice, however, they can be optimized directly with respect to $\mathbf{h}(\boldsymbol{\xi})$ and $\mathbf{g}(\boldsymbol{\eta})$ instead of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. Applications of LVA, storing $\mathbf{h}(\boldsymbol{\xi})$ and $\mathbf{g}(\boldsymbol{\eta})$ as parameters, do not require $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ explicitly. Furthermore, we consider the gradient vectors of the free energy bounds with respect to $\boldsymbol{\theta}(\boldsymbol{\xi})$ and $\boldsymbol{\kappa}(\boldsymbol{\eta})$, which have one-to-one correspondence with $\mathbf{h}(\boldsymbol{\xi})$ and $\mathbf{g}(\boldsymbol{\eta})$, because they provide simple expressions of the gradient vectors. For notational simplicity, we drop the dependencies on $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ and denote as $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$.

The gradient of the upper bound with respect to θ is given by³

$$\begin{aligned}
\nabla_{\theta} \bar{F}(\xi) &= \nabla_{\theta} \left\{ -\log \int \underline{p}(w; \xi) dw \right\} \\
&= - \int \frac{1}{\underline{Z}(\xi)} \frac{\partial \underline{p}(w; \xi)}{\partial \theta} dw \\
&= - \frac{\partial \mathbf{h}(\xi)}{\partial \theta} \int \frac{1}{\underline{Z}(\xi)} \frac{\partial \underline{p}(w; \xi)}{\partial \mathbf{h}(\xi)} dw \\
&= - \frac{\partial \mathbf{h}(\xi)}{\partial \theta} \int \frac{\partial^2 \phi(\mathbf{h}(\xi))}{\partial \mathbf{h} \partial \mathbf{h}^\top} (\mathbf{h}(w) - \mathbf{h}(\xi)) q_\xi(w; \xi) dw \\
&= -(\langle \mathbf{h}(w) \rangle_{q_\xi} - \mathbf{h}(\xi)),
\end{aligned} \tag{5.30}$$

where we have used Eq. (5.22) and the fact that the matrix $\frac{\partial \mathbf{h}(\xi)}{\partial \theta}$, whose (i, j) th entry is $\frac{\partial h_i(\xi)}{\partial \theta_j}$, is the inverse of the Hessian matrix $\frac{\partial^2 \phi(\mathbf{h}(\xi))}{\partial \mathbf{h} \partial \mathbf{h}^\top}$. Similarly, we obtain

$$\nabla_{\eta} \underline{F}(\eta) = \langle \mathbf{g}(w) \rangle_{q_\eta} - \mathbf{g}(\eta). \tag{5.31}$$

Hence, we can utilize gradient methods to minimize $\bar{F}(\xi)$ and maximize $\underline{F}(\eta)$. We can see that when ξ and η are optimized,

$$\mathbf{h}(\xi) = \langle \mathbf{h}(w) \rangle_{q_\xi} \quad \text{and} \quad \mathbf{g}(\eta) = \langle \mathbf{g}(w) \rangle_{q_\eta}$$

hold.

In practice, the variational parameter $\mathbf{h}(\xi)$ is iteratively updated so that $\bar{F}(\xi)$ is monotonically decreased. Recall the argument in Section 2.1.7 where LVA for VB was formulated as the joint minimization of $\bar{F}(\widehat{\lambda}, \xi)$ over the approximate posterior $r(w; \widehat{\lambda})$ and ξ . The free energy bound $\bar{F}(\xi) = -\log \underline{Z}(\xi)$ was obtained as the minimum of $\bar{F}(\widehat{\lambda}, \xi)$, which is attained by (see Eq. (2.32))

$$r(w; \widehat{\lambda}) = q(w; \xi).$$

Let $\mathbf{h}(\xi_o)$ be a current estimate of $\mathbf{h}(\xi)$ and $\widehat{\lambda}_o$ be the variational parameter such that $r(w; \widehat{\lambda}_o) = q(w; \xi_o)$. Then, updating $\mathbf{h}(\xi)$ to $\operatorname{argmin}_{\mathbf{h}(\xi)} \bar{F}(\widehat{\lambda}_o, \xi)$ decreases $\bar{F}(\xi)$ because

$$\bar{F}(\widehat{\lambda}_o, \xi) \geq \bar{F}(\xi)$$

for all ξ and the equality holds when $\xi = \xi_o$. More specifically, it follows for $\widehat{\mathbf{h}}(\widehat{\xi}) = \operatorname{argmin}_{\mathbf{h}(\xi)} \bar{F}(\widehat{\lambda}_o, \xi)$ that

$$\bar{F}(\widehat{\xi}) \leq \bar{F}(\widehat{\lambda}_o, \widehat{\xi}) \leq \bar{F}(\widehat{\lambda}_o, \xi_o) = \bar{F}(\xi_o), \tag{5.32}$$

³ We henceforth use the operator ∇ with the subscript expressing for which variable the gradient is taken. That is, for a function $f(\theta)$, $\nabla_{\theta} f(\theta) = \frac{\partial f(\theta)}{\partial \theta}$ denotes the vector whose i th element is $\frac{\partial f(\theta)}{\partial \theta_i}$.

which means that the bound is improved. This corresponds to the EM algorithm to decrease $\bar{F}(\xi)$ and yields the following specific update rule of $\mathbf{h}(\xi)$:

$$\begin{aligned}\mathbf{h}(\widehat{\xi}) &= \underset{\mathbf{h}(\xi)}{\operatorname{argmin}} \bar{F}(\widehat{\lambda}_o, \widehat{\xi}) \\ &= \underset{\mathbf{h}(\xi)}{\operatorname{argmin}} \left\langle -\log \underline{p}(\mathbf{w}; \xi) \right\rangle_{q_\xi(\mathbf{w}; \xi_o)} \quad (5.33)\end{aligned}$$

$$= \underset{\mathbf{h}(\xi)}{\operatorname{argmin}} \left\langle d_\phi(\mathbf{h}(\mathbf{w}), \mathbf{h}(\xi)) \right\rangle_{q_\xi(\mathbf{w}; \xi_o)} \quad (5.34)$$

$$= \underset{\mathbf{h}(\xi)}{\operatorname{argmin}} d_\phi(\langle \mathbf{h}(\mathbf{w}) \rangle_{q_\xi(\mathbf{w}; \xi_o)}, \mathbf{h}(\xi)) \quad (5.35)$$

$$= \langle \mathbf{h}(\mathbf{w}) \rangle_{q_\xi(\mathbf{w}; \xi_o)}, \quad (5.36)$$

which is summarized as

$$\mathbf{h}(\widehat{\xi}) = \langle \mathbf{h}(\mathbf{w}) \rangle_{q_\xi(\mathbf{w}; \xi_o)}. \quad (5.37)$$

The preceding lines of equations are basically derived by focusing on the terms depending on $\mathbf{h}(\xi)$. Eq. (5.33) follows from the definition of $\bar{F}(\widehat{\lambda}_o, \widehat{\xi})$ by Eq. (2.25). Eq. (5.34) follows from the definition of $\underline{p}(\mathbf{w}; \xi)$ by Eq. (5.15). Eq. (5.35) follows from the definition of the Bregman divergence (5.21) and the linearity of expectation. Eq. (5.36) follows from the nonnegativity of the Bregman divergence. Eqs. (5.34) through (5.36) are equivalent to the fact that the expected Bregman divergence is minimized by the mean (Banerjee et al., 2005).

The update rule (5.37) means that $\mathbf{h}(\xi)$ is updated to the expectation of $\mathbf{h}(\mathbf{w})$ with respect to the approximate posterior. Note here again that if we store $\mathbf{h}(\xi)$, ξ is not explicitly required.

The update rule (5.37) is an iterative substitution of $\mathbf{h}(\xi)$. To maximize the lower-bound $\underline{F}(\boldsymbol{\eta})$ in LVA for EP, such a simple update rule is not applicable in general. Thus, gradient-based optimization methods with the gradient (5.31) are usually used. The Newton–Raphson step to update $\boldsymbol{\kappa}$ is given by

$$\boldsymbol{\kappa}^{\text{new}} = \boldsymbol{\kappa}^{\text{old}} - \left(\nabla_{\boldsymbol{\kappa}}^2 \underline{F}(\boldsymbol{\eta}^{\text{old}}) \right)^{-1} \nabla_{\boldsymbol{\kappa}} \underline{F}(\boldsymbol{\eta}^{\text{old}}), \quad (5.38)$$

where the Hessian matrix is given as follows:

$$\begin{aligned}\nabla_{\boldsymbol{\kappa}}^2 \underline{F}(\boldsymbol{\eta}) &= \frac{\partial^2 \underline{F}(\boldsymbol{\eta})}{\partial \boldsymbol{\kappa} \partial \boldsymbol{\kappa}^\top} \\ &= -\text{Cov}(\mathbf{g}(\mathbf{w})) - \frac{\partial \mathbf{g}(\boldsymbol{\eta})}{\partial \boldsymbol{\kappa}} \quad (5.39)\end{aligned}$$

for the covariance matrix of $\mathbf{g}(\mathbf{w})$,

$$\text{Cov}(\mathbf{g}(\mathbf{w})) = \left\langle \mathbf{g}(\mathbf{w}) \mathbf{g}(\mathbf{w})^\top \right\rangle_{q_\eta} - \langle \mathbf{g}(\mathbf{w}) \rangle_{q_\eta} \langle \mathbf{g}(\mathbf{w}) \rangle_{q_\eta}^\top,$$

and $\frac{\partial \mathbf{g}(\boldsymbol{\eta})}{\partial \boldsymbol{\kappa}} = \left(\frac{\partial^2 \psi(\mathbf{g}(\boldsymbol{\eta}))}{\partial \mathbf{g} \partial \mathbf{g}^\top} \right)^{-1}$ holds in Eq. (5.39).

5.3.3 An Alternative View of VB for Latent Variable Models

In this subsection, we show that the VB learning for latent variable models can be viewed as a special case of LVA, where the log-sum-exp function is used to form the lower-bound of the log-likelihood (Jordan et al., 1999).

Let \mathcal{H} be a vector of latent (unobserved) variables and consider the latent variable model,

$$p(\mathcal{D}, \mathbf{w}) = \sum_{\mathcal{H}} p(\mathcal{D}, \mathcal{H}, \mathbf{w}),$$

where $\sum_{\mathcal{H}}$ denotes the summation over all possible realizations of the latent variables. We have used the notation as if \mathcal{H} were discrete in order to include examples such as GMMs and HMMs, where the likelihood function is given by $p(\mathcal{D}|\mathbf{w}) = \sum_{\mathcal{H}} p(\mathcal{D}, \mathcal{H}|\mathbf{w})$. In the case of a model with continuous latent variables, the summation $\sum_{\mathcal{H}}$ is simply replaced by the integration $\int d\mathcal{H}$. This includes, for example, the hierarchical prior distribution presented in Tipping (2001), where the prior distribution is defined by $p(\mathbf{w}) = \int p(\mathbf{w}|\mathcal{H})p(\mathcal{H})d\mathcal{H}$ with the hyperprior $p(\mathcal{H})$.

The Bayesian posterior distribution of the latent variables and the parameter \mathbf{w} is

$$p(\mathcal{H}, \mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}, \mathcal{H}, \mathbf{w})}{\sum_{\mathcal{H}} \int p(\mathcal{D}, \mathcal{H}, \mathbf{w})d\mathbf{w}},$$

which is intractable when $Z = \sum_{\mathcal{H}} \int p(\mathcal{D}, \mathcal{H}, \mathbf{w})d\mathbf{w}$ requires summation over exponentially many terms as in GMMs and HMMs or the analytically intractable integration. So is the posterior of the parameter $p(\mathbf{w}|\mathcal{D})$.

Let us consider an application of the local variational method for approximating $p(\mathbf{w}|\mathcal{D})$. By the convexity of the function $\log \sum_{\mathcal{H}} \exp(\cdot)$, the log-joint distribution is lower-bounded as follows:

$$\begin{aligned} \log p(\mathcal{D}, \mathbf{w}) &= \log \sum_{\mathcal{H}} \exp\{\log p(\mathcal{D}, \mathcal{H}, \mathbf{w})\} \\ &\geq \log p(\mathcal{D}, \boldsymbol{\xi}) + \sum_{\mathcal{H}} \left(\log \frac{p(\mathcal{D}, \mathcal{H}, \mathbf{w})}{p(\mathcal{D}, \mathcal{H}, \boldsymbol{\xi})} \right) p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi}) \\ &= \log p(\mathcal{D}, \mathbf{w}) - \sum_{\mathcal{H}} p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi}) \log \frac{p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi})}{p(\mathcal{H}|\mathcal{D}, \mathbf{w})}, \end{aligned} \quad (5.40)$$

where $p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi}) = \frac{p(\mathcal{D}, \mathcal{H}, \boldsymbol{\xi})}{\sum_{\mathcal{H}} p(\mathcal{D}, \mathcal{H}, \boldsymbol{\xi})}$. This corresponds to the case where $\phi(\mathbf{h}) = \log \sum_n \exp(h_n)$ and $\mathbf{h}(\mathbf{w})$ is the vector-valued function that consists of the elements $\log p(\mathcal{D}, \mathcal{H}, \mathbf{w})$ for all possible \mathcal{H} . The vector \mathbf{h} is infinite dimensional when \mathcal{H} is continuous. Taking exponentials of the most right-hand side and

left-hand side of Inequality (5.40) leads to Eqs. (5.22) and (5.15) with the Bregman divergence,

$$\begin{aligned} d_\phi(\mathbf{h}(\mathbf{w}), \mathbf{h}(\boldsymbol{\xi})) &= \sum_{\mathcal{H}} p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi}) \log \frac{p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi})}{p(\mathcal{H}|\mathcal{D}, \mathbf{w})} \\ &= \text{KL}(p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi}) \| p(\mathcal{H}|\mathcal{D}, \mathbf{w})). \end{aligned}$$

From Eq. (5.26), we have

$$\begin{aligned} \bar{F}(\boldsymbol{\xi}) &= F^{\text{Bayes}} + \text{KL}(q_\xi(\mathbf{w}; \boldsymbol{\xi}) \| p(\mathbf{w}|\mathcal{D})) + \langle \text{KL}(p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi}) \| p(\mathcal{H}|\mathcal{D}, \mathbf{w})) \rangle_{q_\xi(\mathbf{w}; \boldsymbol{\xi})} \\ &= F^{\text{Bayes}} + \text{KL}(q_\xi(\mathbf{w}; \boldsymbol{\xi}) p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi}) \| p(\mathbf{w}, \mathcal{H}|\mathcal{D})), \end{aligned}$$

which is exactly the free energy of the factorized distribution $q_\xi(\mathbf{w}; \boldsymbol{\xi}) p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi})$. In fact, from Eqs. (5.29) and (5.40), the approximating posterior is given by

$$\begin{aligned} q_\xi(\mathbf{w}; \boldsymbol{\xi}) &\propto \exp \left\{ \sum_{\mathcal{H}} \log p(\mathcal{D}, \mathcal{H}, \mathbf{w}) p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi}) \right\} \\ &= \exp \langle \log p(\mathcal{D}, \mathcal{H}, \mathbf{w}) \rangle_{p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi})}. \end{aligned} \quad (5.41)$$

From Eq. (5.37), the EM update for the variational parameters $\boldsymbol{\xi}$ yields

$$\begin{aligned} \log p(\mathcal{D}, \mathcal{H}, \boldsymbol{\xi}) &= \langle \log p(\mathcal{D}, \mathcal{H}, \mathbf{w}) \rangle_{q_\xi(\mathbf{w}; \boldsymbol{\xi}_0)} \\ &\Rightarrow p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi}) \propto \exp \langle \log p(\mathcal{D}, \mathcal{H}, \mathbf{w}) \rangle_{q_\xi(\mathbf{w}; \boldsymbol{\xi}_0)}. \end{aligned} \quad (5.42)$$

Eqs. (5.41) and (5.42) are exactly the same as the VB algorithm for minimizing the free energy over the factorized distributions, Eqs. (4.2) and (4.3). In this example, we no longer have $\boldsymbol{\xi}$ satisfying Eq. (5.42) in general. However, if the model $p(\mathcal{H}, \mathcal{D}|\mathbf{w})$ and the prior $p(\mathbf{w})$ are included in the exponential family, $\mathbf{h}(\boldsymbol{\xi})$ as well as $p(\mathcal{H}|\mathcal{D}, \boldsymbol{\xi})$ and $q_\xi(\mathbf{w}; \boldsymbol{\xi})$ are expressed by expected sufficient statistics, the number of which is equal to the dimensionality of \mathbf{w} (Beal, 2003). In that case, it is not necessary to obtain $\boldsymbol{\xi}$ explicitly but only to store and update the expected sufficient statistics instead.

Part III

Nonasymptotic Theory

6

Global VB Solution of Fully Observed Matrix Factorization

Variational Bayesian (VB) learning has shown good performance in many applications. However, VB learning sometimes gives a seemingly different posterior and exhibits different sparsity behavior from full Bayesian learning. For example, Figure 6.1 compares the Bayes posterior (left) and the VB posterior (right) of 1×1 matrix factorization. VB posterior tries to approximate a *two-mode* Bayes posterior with a single-mode Gaussian, which results in the zero-mean Gaussian posterior with the VB estimator $\widehat{BA} = 0$. This behavior makes the VB estimator *exactly* sparse as shown in Figure 6.2: thresholding is observed for the VB estimator, while no thresholding is observed for the full Bayesian estimator. Mackay (2001) discussed the sparsity of VB learning as an artifact by showing *inappropriate* model pruning in mixture models. These facts might deprive the justification of VB learning based solely on the fact that it is a tractable approximation to Bayesian learning. Can we clarify the behavior of VB learning, and directly justify its use as an inference method? The nonasymptotic theory, introduced in Part III, gives some answer to this question.

In this chapter, we derive an analytic-form of the *global* VB solution of fully observed matrix factorization (MF). The analytic-form solution allows us to make intuitive discussion on the behavior of VB learning (Chapter 7), and further analysis gives theoretical guarantees of the performance of VB learning (Chapter 8). The analytic-form global solution naturally leads to efficient and reliable algorithms (Chapter 9), which are extended to other similar models (Chapters 10 and 11). Relation to MAP learning and partially Bayesian learning is also theoretically investigated (Chapter 12).

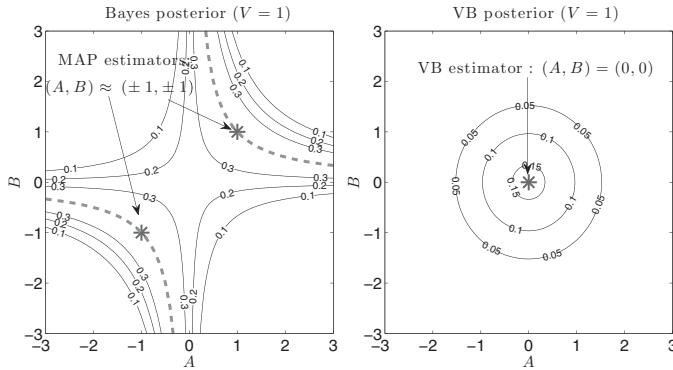


Figure 6.1 The Bayes posterior (left) and the VB posterior (right) of the 1×1 MF model $V = BA + \mathcal{E}$ with almost flat prior, when $V = 1$ is observed (\mathcal{E} is the standard Gaussian noise). VB approximates the Bayes posterior having two modes by an origin-centered Gaussian, which induces sparsity.

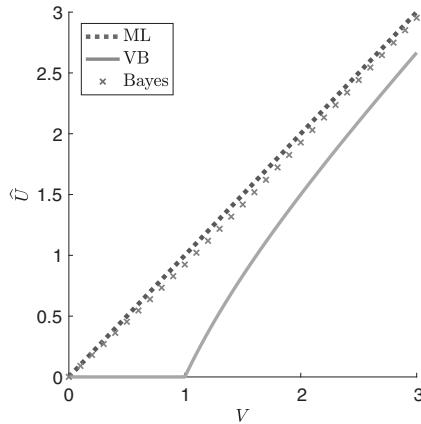


Figure 6.2 Behavior of the estimators of $\widehat{U} = \widehat{BA}$ as a function of the observed value V . The VB estimator is zero when $V \leq 1$, which indicates *exact* sparsity. On the other hand, the Bayesian estimator shows no sign of sparsity. The maximum likelihood estimator, i.e., $\widehat{U} = V$, is shown as a reference.

6.1 Problem Description

We first summarize the MF model and its VB learning algorithm, which was derived in Section 3.1. The likelihood and priors are given as

$$p(V|A, B) \propto \exp\left(-\frac{1}{2\sigma^2} \|V - BA^\top\|_{\text{Fro}}^2\right), \quad (6.1)$$

$$p(A) \propto \exp\left(-\frac{1}{2}\text{tr}(AC_A^{-1}A^\top)\right), \quad (6.2)$$

$$p(\mathbf{B}) \propto \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{B}\mathbf{C}_B^{-1}\mathbf{B}^\top\right)\right), \quad (6.3)$$

where the prior covariances are restricted to be diagonal:

$$\begin{aligned}\mathbf{C}_A &= \mathbf{Diag}(c_{a_1}^2, \dots, c_{a_H}^2), \\ \mathbf{C}_B &= \mathbf{Diag}(c_{b_1}^2, \dots, c_{b_H}^2),\end{aligned}$$

for $c_{a_h}, c_{b_h} > 0, h = 1, \dots, H$. Without loss of generality, we assume that the diagonal entries of the product $\mathbf{C}_A \mathbf{C}_B$ are arranged in nonincreasing order, i.e., $c_{a_h} c_{b_h} \geq c_{a_{h'}} c_{b_{h'}}$ for any pair $h < h'$. We assume that

$$L \leq M. \quad (6.4)$$

If $L > M$, we may simply redefine the transpose \mathbf{V}^\top as \mathbf{V} so that $L \leq M$ holds. Therefore, the assumption (6.4) does not impose any restriction.

We solve the following VB learning problem:

$$\widehat{\mathbf{r}} = \underset{r}{\operatorname{argmin}} F(r) \quad \text{s.t.} \quad r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}), \quad (6.5)$$

where the objective function to be minimized is the free energy:

$$F = \left\langle \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{p(\mathbf{V}|\mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}.$$

The solution to the problem (6.5) is in the following form:

$$r_A(\mathbf{A}) = \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \mathbf{I}_M \otimes \widehat{\Sigma}_A) \propto \exp\left(-\frac{\text{tr}\left((\mathbf{A} - \widehat{\mathbf{A}})\widehat{\Sigma}_A^{-1}(\mathbf{A} - \widehat{\mathbf{A}})^\top\right)}{2}\right), \quad (6.6)$$

$$r_B(\mathbf{B}) = \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \mathbf{I}_L \otimes \widehat{\Sigma}_B) \propto \exp\left(-\frac{\text{tr}\left((\mathbf{B} - \widehat{\mathbf{B}})\widehat{\Sigma}_B^{-1}(\mathbf{B} - \widehat{\mathbf{B}})^\top\right)}{2}\right). \quad (6.7)$$

With the variational parameters $\widehat{\mathbf{A}}, \widehat{\Sigma}_A, \widehat{\mathbf{B}}, \widehat{\Sigma}_B$, the free energy can be explicitly written as

$$\begin{aligned}2F &= LM \log(2\pi\sigma^2) + \frac{\|\mathbf{V} - \widehat{\mathbf{B}}\mathbf{A}^\top\|_{\text{Fro}}^2}{\sigma^2} + M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + L \log \frac{\det(\mathbf{C}_B)}{\det(\widehat{\Sigma}_B)} \\ &\quad - (L+M)H + \text{tr}\left\{\mathbf{C}_A^{-1}\left(\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\Sigma}_A\right)\right\} + \text{tr}\left\{\mathbf{C}_B^{-1}\left(\widehat{\mathbf{B}}^\top\widehat{\mathbf{B}} + L\widehat{\Sigma}_B\right)\right\} \\ &\quad + \sigma^{-2} \text{tr}\left\{-\widehat{\mathbf{A}}^\top\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top\widehat{\mathbf{B}} + \left(\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} + M\widehat{\Sigma}_A\right)\left(\widehat{\mathbf{B}}^\top\widehat{\mathbf{B}} + L\widehat{\Sigma}_B\right)\right\}.\end{aligned} \quad (6.8)$$

The stationary conditions for the variational parameters are given by

$$\widehat{\mathbf{A}} = \sigma^{-2} \mathbf{V}^\top \widetilde{\mathbf{B} \Sigma}_A, \quad (6.9)$$

$$\widehat{\Sigma}_A = \sigma^2 \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B + \sigma^2 \mathbf{C}_A^{-1} \right)^{-1}, \quad (6.10)$$

$$\widehat{\mathbf{B}} = \sigma^{-2} \mathbf{V} \widehat{\mathbf{A} \Sigma}_B, \quad (6.11)$$

$$\widehat{\Sigma}_B = \sigma^2 \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A + \sigma^2 \mathbf{C}_B^{-1} \right)^{-1}. \quad (6.12)$$

In the subsequent sections, we derive the global solution to the problem (6.5) in an analytic form, which was obtained in Nakajima et al. (2013a, 2015).

6.2 Conditions for VB Solutions

With the explicit form (6.8) of the free energy, the VB learning problem (6.5) can be written as a minimization problem with respect to a *finite* number of variables:

$$\begin{aligned} \text{Given } \mathbf{C}_A, \mathbf{C}_B \in \mathbb{D}_{++}^H, \quad \sigma^2 \in \mathbb{R}_{++}, \\ \min_{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B} F \end{aligned} \quad (6.13)$$

$$\text{s.t. } \widehat{\mathbf{A}} \in \mathbb{R}^{M \times H}, \widehat{\mathbf{B}} \in \mathbb{R}^{L \times H}, \quad \widehat{\Sigma}_A, \widehat{\Sigma}_B \in \mathbb{S}_{++}^H. \quad (6.14)$$

We can easily show that the solution is a stationary point of the free energy.

Lemma 6.1 *Any local solution of the problem (6.13) is a stationary point of the free energy (6.8).*

Proof Since

$$\left\| \mathbf{V} - \widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top \right\|_{\text{Fro}}^2 \geq 0,$$

and

$$\begin{aligned} & \text{tr} \left\{ -\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B \right) \right\} \\ &= \text{tr} \left\{ L \widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\Sigma}_B + M \widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} \widehat{\Sigma}_A + LM \widehat{\Sigma}_A \widehat{\Sigma}_B \right\} \geq 0, \end{aligned}$$

the free energy (6.8) is lower-bounded as

$$\begin{aligned} 2F &\geq -M \log \det(\widehat{\Sigma}_A) - L \log \det(\widehat{\Sigma}_B) \\ &+ \text{tr} \left\{ \mathbf{C}_A^{-1} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) \right\} + \text{tr} \left\{ \mathbf{C}_B^{-1} \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B \right) \right\} + \tau, \end{aligned} \quad (6.15)$$

where τ is a finite constant. The right-hand side of Eq. (6.15) diverges to $+\infty$ if any entry of $\widehat{\mathbf{A}}$ or $\widehat{\mathbf{B}}$ goes to $+\infty$ or $-\infty$. Also it diverges if any eigenvalue of $\widehat{\Sigma}_A$ or $\widehat{\Sigma}_B$ goes to $+0$ or ∞ . This implies that no local solution exists on the boundary of (the closure of) the domain (6.14). Since the free energy is differentiable in the domain (6.14), any local minimizer is a stationary point.

For any observed matrix \mathbf{V} , the free energy (6.8) can be finite, for example, at $\widehat{\mathbf{A}} = \mathbf{0}_{M,H}$, $\widehat{\mathbf{B}} = \mathbf{0}_{L,H}$, and $\widehat{\Sigma}_A = \widehat{\Sigma}_B = \mathbf{I}_H$, where $\mathbf{0}_{D_1,D_2}$ denotes the $D_1 \times D_2$ matrix with all the entries equal to zero. Therefore, at least one minimizer always exists, which completes the proof of Lemma 6.1. \square

Lemma 6.1 implies that the stationary conditions (6.9) through (6.12) are satisfied for any solution. Accordingly, we can obtain the global solution by finding all points that satisfy the stationary conditions. However, the condition involves $O(MH)$ unknowns, and therefore finding all such candidate points seems hard. The first step to tackle this problem is to find hidden separability, which enables us to decompose the problem so that each problem involves only $O(1)$ unknowns.

6.3 Irrelevant Degrees of Freedom

The most of the terms in the free energy (6.8) have symmetry, i.e., they are invariant with respect to the coordinate change shown in Eqs. (6.16) and (6.17). Assume that $(\mathbf{A}^*, \mathbf{B}^*, \Sigma_A^*, \Sigma_B^*)$ is a global solution of the VB problem (6.13), and let $F^* = F(\mathbf{A}^*, \mathbf{B}^*, \Sigma_A^*, \Sigma_B^*)$ be the minimum free energy. Consider the following rotation of the coordinate system for an arbitrary orthogonal matrix $\boldsymbol{\Omega} \in \mathbb{R}^{H \times H}$:

$$\widehat{\mathbf{A}} = \mathbf{A}^* \boldsymbol{\Omega}^\top, \quad \widehat{\Sigma}_A = \boldsymbol{\Omega} \Sigma_A^* \boldsymbol{\Omega}^\top, \quad (6.16)$$

$$\widehat{\mathbf{B}} = \mathbf{B}^* \boldsymbol{\Omega}^\top, \quad \widehat{\Sigma}_B = \boldsymbol{\Omega} \Sigma_B^* \boldsymbol{\Omega}^\top. \quad (6.17)$$

We can easily confirm that the terms in Eq. (6.8) except the sixth and the seventh terms are invariant with respect to the rotation, and the free energy can be written as a function of $\boldsymbol{\Omega}$ as follows:

$$2F(\boldsymbol{\Omega}) = \text{tr} \left\{ \mathbf{C}_A^{-1} \boldsymbol{\Omega} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) \boldsymbol{\Omega}^\top \right\} + \text{tr} \left\{ \mathbf{C}_B^{-1} \boldsymbol{\Omega} \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B \right) \boldsymbol{\Omega}^\top \right\} + \text{const.}$$

To find the irrelevant degrees of freedom, we consider *skewed* rotations that only affect a single term in Eq. (6.8).

Consider the following transform:

$$\widehat{\mathbf{A}} = \mathbf{A}^* \mathbf{C}_A^{-1/2} \boldsymbol{\Omega}^\top \mathbf{C}_A^{1/2}, \quad \Sigma_A = \mathbf{C}_A^{1/2} \boldsymbol{\Omega} \mathbf{C}_A^{-1/2} \Sigma_A^* \mathbf{C}_A^{-1/2} \boldsymbol{\Omega}^\top \mathbf{C}_A^{1/2}, \quad (6.18)$$

$$\widehat{\mathbf{B}} = \mathbf{B}^* \mathbf{C}_A^{1/2} \boldsymbol{\Omega}^\top \mathbf{C}_A^{-1/2}, \quad \boldsymbol{\Sigma}_{\mathbf{B}} = \mathbf{C}_A^{-1/2} \boldsymbol{\Omega} \mathbf{C}_A^{1/2} \boldsymbol{\Sigma}_{\mathbf{B}}^* \mathbf{C}_A^{1/2} \boldsymbol{\Omega}^\top \mathbf{C}_A^{-1/2}. \quad (6.19)$$

Then, the free energy can be written as

$$2F(\boldsymbol{\Omega}) = \text{tr}\{\boldsymbol{\Gamma} \boldsymbol{\Omega} \boldsymbol{\Phi} \boldsymbol{\Omega}^\top\} + \text{const.}, \quad (6.20)$$

where

$$\begin{aligned} \boldsymbol{\Gamma} &= \mathbf{C}_A^{-1} \mathbf{C}_B^{-1}, \\ \boldsymbol{\Phi} &= \mathbf{C}_A^{1/2} (\mathbf{B}^{*\top} \mathbf{B}^* + L \boldsymbol{\Sigma}_{\mathbf{B}}^*) \mathbf{C}_A^{1/2}. \end{aligned}$$

By assumption, $\boldsymbol{\Omega} = \mathbf{I}_H$ is a minimizer of Eq. (6.20), i.e., $F(\mathbf{I}_H) = F^*$. Now we can use the following lemma:

Lemma 6.2 *Let $\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Phi} \in \mathbb{R}^{H \times H}$ be a nondegenerate diagonal matrix, an orthogonal matrix, and a symmetric matrix, respectively. Let $\{\boldsymbol{\Lambda}^{(k)}, \boldsymbol{\Lambda}'^{(k)} \in \mathbb{R}^{H \times H}; k = 1, \dots, K\}$ be arbitrary diagonal matrices. If a function*

$$G(\boldsymbol{\Omega}) = \text{tr}\left\{\boldsymbol{\Gamma} \boldsymbol{\Omega} \boldsymbol{\Phi} \boldsymbol{\Omega}^\top + \sum_{k=1}^K \boldsymbol{\Lambda}^{(k)} \boldsymbol{\Omega} \boldsymbol{\Lambda}'^{(k)} \boldsymbol{\Omega}^\top\right\} \quad (6.21)$$

is minimized (as a function of $\boldsymbol{\Omega}$, given $\boldsymbol{\Gamma}, \boldsymbol{\Phi}, \{\boldsymbol{\Lambda}^{(k)}, \boldsymbol{\Lambda}'^{(k)}\}$) at $\boldsymbol{\Omega} = \mathbf{I}_H$, then $\boldsymbol{\Phi}$ is diagonal. Here, K can be any natural number including $K = 0$ (when only the first term exists).

Proof Let

$$\boldsymbol{\Phi} = \boldsymbol{\Omega}' \boldsymbol{\Gamma}' \boldsymbol{\Omega}'^\top \quad (6.22)$$

be the eigenvalue decomposition of $\boldsymbol{\Phi}$. Let $\boldsymbol{\gamma}, \boldsymbol{\gamma}', \{\boldsymbol{\lambda}^{(k)}\}, \{\boldsymbol{\lambda}'^{(k)}\}$ be the vectors consist of the diagonal entries of $\boldsymbol{\Gamma}, \boldsymbol{\Gamma}', \{\boldsymbol{\Lambda}^{(k)}\}, \{\boldsymbol{\Lambda}'^{(k)}\}$, respectively, i.e.,

$$\boldsymbol{\Gamma} = \text{Diag}(\boldsymbol{\gamma}), \quad \boldsymbol{\Gamma}' = \text{Diag}(\boldsymbol{\gamma}'), \quad \boldsymbol{\Lambda}^{(k)} = \text{Diag}(\boldsymbol{\lambda}^{(k)}), \quad \boldsymbol{\Lambda}'^{(k)} = \text{Diag}(\boldsymbol{\lambda}'^{(k)}).$$

Then, Eq. (6.21) can be written as

$$G(\boldsymbol{\Omega}) = \text{tr}\left\{\boldsymbol{\Gamma} \boldsymbol{\Omega} \boldsymbol{\Phi} \boldsymbol{\Omega}^\top + \sum_{k=1}^K \boldsymbol{\Lambda}^{(k)} \boldsymbol{\Omega} \boldsymbol{\Lambda}'^{(k)} \boldsymbol{\Omega}^\top\right\} = \boldsymbol{\gamma}^\top \boldsymbol{Q} \boldsymbol{\gamma}' + \sum_{k=1}^K \boldsymbol{\lambda}^{(k)\top} \boldsymbol{R} \boldsymbol{\lambda}'^{(k)}, \quad (6.23)$$

where

$$\boldsymbol{Q} = (\boldsymbol{\Omega} \boldsymbol{\Omega}') \odot (\boldsymbol{\Omega} \boldsymbol{\Omega}'), \quad \boldsymbol{R} = \boldsymbol{\Omega} \odot \boldsymbol{\Omega}.$$

Here, \odot denotes the *Hadamard product*. Since \boldsymbol{Q} as well as \boldsymbol{R} is the Hadamard square of an orthogonal matrix, it is *doubly stochastic* (i.e., any of the columns and the rows sums up to one) (Marshall et al., 2009). Therefore, it can be seen that \boldsymbol{Q} reassigns the components of $\boldsymbol{\gamma}$ to those of $\boldsymbol{\gamma}'$ when calculating

the elementwise product in the first term of Eq. (6.23). The same applies to \mathbf{R} and $\{\lambda^{(k)}, \boldsymbol{\lambda}^{(k)}\}$ in the second term. Naturally, rearranging the components of $\boldsymbol{\gamma}$ in nondecreasing order and the components of $\boldsymbol{\gamma}'$ in nonincreasing order minimizes $\boldsymbol{\gamma}^\top \mathbf{Q} \boldsymbol{\gamma}'$ (Ruhe, 1970; Marshall et al., 2009).

Using the expression (6.23) with \mathbf{Q} and \mathbf{R} , we will prove that $\boldsymbol{\Phi}$ is diagonal if $\boldsymbol{\Omega} = \mathbf{I}_H$ minimizes Eq. (6.23). Let us consider a bilateral perturbation $\boldsymbol{\Omega} = \boldsymbol{\Delta}$ such that the 2×2 matrix $\boldsymbol{\Delta}_{(h,h')}$ consisting of the h th and the h' th columns and rows form an 2×2 orthogonal matrix

$$\boldsymbol{\Delta}_{(h,h')} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix},$$

and the remaining entries coincide with those of the identity matrix. Then, the elements of \mathbf{Q} become

$$Q_{i,j} = \begin{cases} (\Omega'_{h,j} \cos \theta - \Omega'_{h',j} \sin \theta)^2 & \text{if } i = h, \\ (\Omega'_{h,j} \sin \theta + \Omega'_{h',j} \cos \theta)^2 & \text{if } i = h', \\ \Omega'^2_{i,j} & \text{otherwise,} \end{cases}$$

and Eq. (6.23) can be written as a function of θ :

$$\begin{aligned} G(\theta) &= \sum_{j=1}^H \left\{ \gamma_h (\Omega'_{h,j} \cos \theta - \Omega'_{h',j} \sin \theta)^2 + \gamma_{h'} (\Omega'_{h,j} \sin \theta + \Omega'_{h',j} \cos \theta)^2 \right\} \gamma'_j \\ &\quad + \sum_{k=1}^K \begin{pmatrix} \lambda_h^{(k)} & \lambda_{h'}^{(k)} \end{pmatrix} \begin{pmatrix} \cos^2 \theta & \sin^2 \theta \\ \sin^2 \theta & \cos^2 \theta \end{pmatrix} \begin{pmatrix} \lambda_h^{(k)} \\ \lambda_{h'}^{(k)} \end{pmatrix} + \text{const.} \end{aligned} \quad (6.24)$$

Since Eq. (6.24) is differentiable at $\theta = 0$, our assumption that Eq. (6.23) is minimized when $\boldsymbol{\Omega} = \mathbf{I}_H$ requires that $\theta = 0$ is a stationary point of Eq. (6.24) for any $h \neq h'$. Therefore, it holds that

$$\begin{aligned} 0 &= \frac{\partial G}{\partial \theta} \Big|_{\theta=0} = 2 \sum_j \left\{ \gamma_h (\Omega'_{h,j} \cos \theta - \Omega'_{h',j} \sin \theta) (-\Omega'_{h,j} \sin \theta - \Omega'_{h',j} \cos \theta) \right. \\ &\quad \left. + \gamma_{h'} (\Omega'_{h,j} \sin \theta + \Omega'_{h',j} \cos \theta) (\Omega'_{h,j} \cos \theta - \Omega'_{h',j} \sin \theta) \right\} \gamma'_j \\ &= 2 (\gamma_{h'} - \gamma_h) \sum_j \Omega'_{h,j} \gamma'_j \Omega'_{h',j} = 2 (\gamma_{h'} - \gamma_h) \Phi_{h,h'}. \end{aligned} \quad (6.25)$$

In the last equality, we used Eq. (6.22). Since we assumed that $\boldsymbol{\Gamma}$ is nondegenerate ($\gamma_h \neq \gamma_{h'}$ for $h \neq h'$), Eq. (6.25) implies that $\boldsymbol{\Phi}$ is diagonal, which completes the proof of Lemma 6.2. \square

Assume for simplicity that $\boldsymbol{\Gamma} = \mathbf{C}_A^{-1} \mathbf{C}_B^{-1}$ is nondegenerate, i.e., no pair of diagonal entries coincide, in Eq. (6.20). Then, since Eq. (6.20) is minimized

at $\boldsymbol{\Omega} = \mathbf{I}_H$, Lemma 6.2 implies that $\boldsymbol{\Phi} = \mathbf{C}_A^{1/2} (\mathbf{B}^{*\top} \mathbf{B}^* + L\boldsymbol{\Sigma}_B^*) \mathbf{C}_A^{1/2}$ is diagonal. This means that $\mathbf{B}^{*\top} \mathbf{B}^* + L\boldsymbol{\Sigma}_B^*$ is diagonal. Thus, the stationary condition (6.10) implies that $\boldsymbol{\Sigma}_A^*$ is diagonal. Similarly, we can find that $\boldsymbol{\Sigma}_B^*$ is diaognal, if $\boldsymbol{\Gamma} = \mathbf{C}_A^{-1} \mathbf{C}_B^{-1}$ is nondegenerate.

To generalize the preceding discussion to degenerate cases, we need to consider an *equivalent solution*, defined as follows:

Definition 6.3 (Equivalent solutions) We say that two points $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\Sigma}}_A, \widehat{\boldsymbol{\Sigma}}_B)$ and $(\widehat{\mathbf{A}}', \widehat{\mathbf{B}}', \widehat{\boldsymbol{\Sigma}}'_A, \widehat{\boldsymbol{\Sigma}}'_B)$ are *equivalent* if both give the same free energy and the same mean prediction, i.e.,

$$F(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\Sigma}}_A, \widehat{\boldsymbol{\Sigma}}_B) = F(\widehat{\mathbf{A}}', \widehat{\mathbf{B}}', \widehat{\boldsymbol{\Sigma}}'_A, \widehat{\boldsymbol{\Sigma}}'_B) \quad \text{and} \quad \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top = \widehat{\mathbf{B}}'\widehat{\mathbf{A}}'^\top.$$

With this definition, we can obtain the following theorem (its proof is given in the next section):

Theorem 6.4 When $\mathbf{C}_A \mathbf{C}_B$ is nondegenerate (i.e., $c_{a_h} c_{b_h} > c_{a_{h'}} c_{b_{h'}}$ for any pair $h < h'$), any solution of the problem (6.13) has diagonal $\widehat{\boldsymbol{\Sigma}}_A$ and $\widehat{\boldsymbol{\Sigma}}_B$. When $\mathbf{C}_A \mathbf{C}_B$ is degenerate, any solution has an equivalent solution with diagonal $\widehat{\boldsymbol{\Sigma}}_A$ and $\widehat{\boldsymbol{\Sigma}}_B$.

The result that the solution has diagonal $\widehat{\boldsymbol{\Sigma}}_A$ and $\widehat{\boldsymbol{\Sigma}}_B$ would be natural because we assumed the independent Gaussian priors on \mathbf{A} and \mathbf{B} : the fact that any \mathbf{V} can be decomposed into orthogonal singular components may imply that the observation \mathbf{V} cannot convey any preference for singular-componentwise correlation. Note, however, that Theorem 6.4 does not necessarily hold when the observed matrix has missing entries.

Obviously, any VB solution (a solution of the problem (6.13)) with diagonal covariances can be found by solving the following problem:

$$\text{Given } \mathbf{C}_A, \mathbf{C}_B \in \mathbb{D}_{++}^H, \quad \sigma^2 \in \mathbb{R}_{++}, \quad (6.26)$$

$$\min_{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\Sigma}}_A, \widehat{\boldsymbol{\Sigma}}_B} F \quad (6.26)$$

$$\text{s.t. } \widehat{\mathbf{A}} \in \mathbb{R}^{M \times H}, \widehat{\mathbf{B}} \in \mathbb{R}^{L \times H}, \quad \widehat{\boldsymbol{\Sigma}}_A, \widehat{\boldsymbol{\Sigma}}_B \in \mathbb{D}_{++}^H, \quad (6.27)$$

which is equivalent to solving the SimpleVB learning problem (3.30) with columnwise independence, introduced in Section 3.1.1. Theorem 6.4 states that, if $\mathbf{C}_A \mathbf{C}_B$ is nondegenerate, the set of VB solutions and the set of SimpleVB solutions are identical. When $\mathbf{C}_A \mathbf{C}_B$ is degenerate, the set of VB solutions is the union of the set of SimpleVB solutions and the set of their *equivalent* solutions with nondiagonal covariances. Actually, any VB solution can be obtained by rotating its *equivalent* SimpleVB solution (VB solution with diagonal covariances) (see Section 6.4.4). In practice, it is however sufficient

to focus on the SimpleVB solutions, since *equivalent* solutions share the same free energy F and the same mean prediction $\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top$. In this sense, we can conclude that the stronger *columnwise* independence constraint (3.29) does not degrade approximation accuracy, and the VB solution under the *matrixwise* independence (3.4) *essentially* agrees with the SimpleVB solution.

6.4 Proof of Theorem 6.4

In this section, we prove Theorem 6.4 by considering the following three cases separately:

Case 1 When no pair of diagonal entries of $\mathbf{C}_A \mathbf{C}_B$ coincide.

Case 2 When all diagonal entries of $\mathbf{C}_A \mathbf{C}_B$ coincide.

Case 3 When (not all but) some pairs of diagonal entries of $\mathbf{C}_A \mathbf{C}_B$ coincide.

We will prove that, in Case 1, $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$ are diagonal for any solution $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B)$, and that, in other cases, any solution has its *equivalent* solution with diagonal $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$.

Remember our assumption that the diagonal entries $\{c_{ah} c_{bh}\}$ of $\mathbf{C}_A \mathbf{C}_B$ are arranged in nonincreasing order.

6.4.1 Proof for Case 1

Here, we consider the case where $c_{ah} c_{bh} > c_{a'h'} c_{b'h'}$ for any pair $h < h'$.

Assume that $(\mathbf{A}^*, \mathbf{B}^*, \Sigma_A^*, \Sigma_B^*)$ is a minimizer of the free energy (6.8), and consider the following variation defined with an arbitrary $H \times H$ orthogonal matrix $\boldsymbol{\Omega}$:

$$\widehat{\mathbf{A}} = \mathbf{A}^* \mathbf{C}_A^{-1/2} \boldsymbol{\Omega}^\top \mathbf{C}_A^{1/2}, \quad (6.28)$$

$$\widehat{\mathbf{B}} = \mathbf{B}^* \mathbf{C}_A^{1/2} \boldsymbol{\Omega}^\top \mathbf{C}_A^{-1/2}, \quad (6.29)$$

$$\widehat{\Sigma}_A = \mathbf{C}_A^{1/2} \boldsymbol{\Omega} \mathbf{C}_A^{-1/2} \Sigma_A^* \mathbf{C}_A^{-1/2} \boldsymbol{\Omega}^\top \mathbf{C}_A^{1/2}, \quad (6.30)$$

$$\widehat{\Sigma}_B = \mathbf{C}_A^{-1/2} \boldsymbol{\Omega} \mathbf{C}_A^{1/2} \Sigma_B^* \mathbf{C}_A^{1/2} \boldsymbol{\Omega}^\top \mathbf{C}_A^{-1/2}. \quad (6.31)$$

Note that this variation does not change $\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top$, and it holds that $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B) = (\mathbf{A}^*, \mathbf{B}^*, \Sigma_A^*, \Sigma_B^*)$ for $\boldsymbol{\Omega} = \mathbf{I}_H$. Then, the free energy (6.8)

can be written as a function of Ω :

$$F(\Omega) = \frac{1}{2} \text{tr} \left\{ C_A^{-1} C_B^{-1} \Omega C_A^{1/2} (B^{*\top} B^* + L \Sigma_B^*) C_A^{1/2} \Omega^\top \right\} + \text{const.} \quad (6.32)$$

We define

$$\Phi = C_A^{1/2} (B^{*\top} B^* + L \Sigma_B^*) C_A^{1/2},$$

and rewrite Eq. (6.32) as

$$F(\Omega) = \frac{1}{2} \text{tr} \left\{ C_A^{-1} C_B^{-1} \Omega \Phi \Omega^\top \right\} + \text{const.} \quad (6.33)$$

The assumption that $(A^*, B^*, \Sigma_A^*, \Sigma_B^*)$ is a minimizer requires that Eq. (6.33) is minimized when $\Omega = I_H$. Then, Lemma 6.2 (for $K = 0$) implies that Φ is diagonal. Therefore,

$$C_A^{-1/2} \Phi C_A^{-1/2} (= \Phi C_A^{-1}) = B^{*\top} B^* + L \Sigma_B^*$$

is also diagonal. Consequently, Eq. (6.10) implies that Σ_A^* is diagonal.

Next, consider the following variation defined with an arbitrary $H \times H$ orthogonal matrix Ω' :

$$\begin{aligned} \widehat{A} &= A^* C_B^{1/2} \Omega'^\top C_B^{-1/2}, \\ \widehat{B} &= B^* C_B^{-1/2} \Omega'^\top C_B^{1/2}, \\ \widehat{\Sigma}_A &= C_B^{-1/2} \Omega' C_B^{1/2} \Sigma_A^* C_B^{1/2} \Omega'^\top C_B^{-1/2}, \\ \widehat{\Sigma}_B &= C_B^{1/2} \Omega' C_B^{-1/2} \Sigma_B^* C_B^{-1/2} \Omega'^\top C_B^{1/2}. \end{aligned}$$

Then, the free energy as a function of Ω' is given by

$$F(\Omega') = \frac{1}{2} \text{tr} \left\{ C_A^{-1} C_B^{-1} \Omega' C_B^{1/2} (A^{*\top} A^* + M \Sigma_A^*) C_B^{1/2} \Omega'^\top \right\} + \text{const.}$$

From this, we can similarly prove that Σ_B^* is also diagonal, which completes the proof for Case 1. \square

6.4.2 Proof for Case 2

Here, we consider the case where $C_A C_B = c^2 I_H$ for some $c^2 \in \mathbb{R}_{++}$. In this case, there exist solutions with nondiagonal covariances. However, for any (or each) of those nondiagonal solutions, the equivalent class to which the (nondiagonal) solution belongs contains a solution with diagonal covariances.

We can easily show that the free energy (6.8) is invariant with respect to Ω under the transformation (6.28) through (6.31). This arbitrariness forms an *equivalent* class of solutions. Since there exists Ω that diagonalizes any given Σ_A^* through Eq. (6.30), each *equivalent* class involves a solution with diagonal

$\widehat{\Sigma}_A$. In the following, we will prove that any solution with diagonal $\widehat{\Sigma}_A$ has diagonal $\widehat{\Sigma}_B$.

Assume that $(\mathbf{A}^*, \mathbf{B}^*, \boldsymbol{\Sigma}_A^*, \boldsymbol{\Sigma}_B^*)$ is a solution with diagonal $\boldsymbol{\Sigma}_A^*$, and consider the following variation defined with an arbitrary $H \times H$ orthogonal matrix $\boldsymbol{\Omega}$:

$$\begin{aligned}\widehat{\mathbf{A}} &= \mathbf{A}^* \mathbf{C}_A^{-1/2} \boldsymbol{\Gamma}^{-1/2} \boldsymbol{\Omega}^\top \boldsymbol{\Gamma}^{1/2} \mathbf{C}_A^{1/2}, \\ \widehat{\mathbf{B}} &= \mathbf{B}^* \mathbf{C}_A^{1/2} \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Omega}^\top \boldsymbol{\Gamma}^{-1/2} \mathbf{C}_A^{-1/2}, \\ \widehat{\boldsymbol{\Sigma}}_A &= \mathbf{C}_A^{1/2} \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Omega} \boldsymbol{\Gamma}^{-1/2} \mathbf{C}_A^{-1/2} \boldsymbol{\Sigma}_A^* \mathbf{C}_A^{-1/2} \boldsymbol{\Gamma}^{-1/2} \boldsymbol{\Omega}^\top \boldsymbol{\Gamma}^{1/2} \mathbf{C}_A^{1/2}, \\ \widehat{\boldsymbol{\Sigma}}_B &= \mathbf{C}_A^{-1/2} \boldsymbol{\Gamma}^{-1/2} \boldsymbol{\Omega} \boldsymbol{\Gamma}^{1/2} \mathbf{C}_A^{1/2} \boldsymbol{\Sigma}_B^* \mathbf{C}_A^{1/2} \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Omega}^\top \boldsymbol{\Gamma}^{-1/2} \mathbf{C}_A^{-1/2}.\end{aligned}$$

Here, $\boldsymbol{\Gamma} = \text{Diag}(\gamma_1, \dots, \gamma_H)$ is an arbitrary nondegenerate ($\gamma_h \neq \gamma_{h'}$ for $h \neq h'$) positive-definite diagonal matrix. Then, the free energy can be written as a function of $\boldsymbol{\Omega}$:

$$\begin{aligned}F(\boldsymbol{\Omega}) &= \frac{1}{2} \text{tr} \left\{ \boldsymbol{\Gamma} \boldsymbol{\Omega} \boldsymbol{\Gamma}^{-1/2} \mathbf{C}_A^{-1/2} \left(\mathbf{A}^{*\top} \mathbf{A}^* + M \boldsymbol{\Sigma}_A^* \right) \mathbf{C}_A^{-1/2} \boldsymbol{\Gamma}^{-1/2} \boldsymbol{\Omega}^\top \right. \\ &\quad \left. + c^{-2} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Omega} \boldsymbol{\Gamma}^{1/2} \mathbf{C}_A^{1/2} \left(\mathbf{B}^{*\top} \mathbf{B}^* + L \boldsymbol{\Sigma}_B^* \right) \mathbf{C}_A^{1/2} \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Omega}^\top \right\}. \quad (6.34)\end{aligned}$$

We define

$$\begin{aligned}\boldsymbol{\Phi}_A &= \boldsymbol{\Gamma}^{-1/2} \mathbf{C}_A^{-1/2} \left(\mathbf{A}^{*\top} \mathbf{A}^* + M \boldsymbol{\Sigma}_A^* \right) \mathbf{C}_A^{-1/2} \boldsymbol{\Gamma}^{-1/2}, \\ \boldsymbol{\Phi}_B &= c^{-2} \boldsymbol{\Gamma}^{1/2} \mathbf{C}_A^{1/2} \left(\mathbf{B}^{*\top} \mathbf{B}^* + L \boldsymbol{\Sigma}_B^* \right) \mathbf{C}_A^{1/2} \boldsymbol{\Gamma}^{1/2},\end{aligned}$$

and rewrite Eq. (6.34) as

$$F(\boldsymbol{\Omega}) = \frac{1}{2} \text{tr} \left\{ \boldsymbol{\Gamma} \boldsymbol{\Omega} \boldsymbol{\Phi}_A \boldsymbol{\Omega}^\top + \boldsymbol{\Gamma}^{-1} \boldsymbol{\Omega} \boldsymbol{\Phi}_B \boldsymbol{\Omega}^\top \right\}. \quad (6.35)$$

Since $\boldsymbol{\Sigma}_A^*$ is diagonal, Eq. (6.10) implies that $\boldsymbol{\Phi}_B$ is diagonal. The assumption that $(\mathbf{A}^*, \mathbf{B}^*, \boldsymbol{\Sigma}_A^*, \boldsymbol{\Sigma}_B^*)$ is a solution requires that Eq. (6.35) is minimized when $\boldsymbol{\Omega} = \mathbf{I}_H$. Accordingly, Lemma 6.2 implies that $\boldsymbol{\Phi}_A$ is diagonal. Consequently, Eq. (6.12) implies that $\boldsymbol{\Sigma}_B^*$ is diagonal.

Thus, we have proved that any solution has its *equivalent* solution with diagonal covariances, which completes the proof for Case 2. \square

6.4.3 Proof for Case 3

Finally, we consider the case where $c_{ah}c_{bh} = c_{ah'}c_{bh'}$ for (not all but) some pairs $h \neq h'$. First, in the same way as Case 1, we can prove that $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$ are block diagonal where the blocks correspond to the groups sharing the same $c_{ah}c_{bh}$. Next, we can apply the proof for Case 2 to each block, and show that any solution has its *equivalent* solution with diagonal $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$. Combining these results completes the proof of Theorem 6.4. \square

6.4.4 General Expression

In summary, for any minimizer of Eq. (6.8), the covariances can be written in the following form:

$$\widehat{\Sigma}_A = \mathbf{C}_A^{1/2} \boldsymbol{\Theta} \mathbf{C}_A^{-1/2} \boldsymbol{\Gamma}_{\widehat{\Sigma}_A} \mathbf{C}_A^{-1/2} \boldsymbol{\Theta}^\top \mathbf{C}_A^{1/2} (= \mathbf{C}_B^{-1/2} \boldsymbol{\Theta} \mathbf{C}_B^{1/2} \boldsymbol{\Gamma}_{\widehat{\Sigma}_A} \mathbf{C}_B^{1/2} \boldsymbol{\Theta}^\top \mathbf{C}_B^{-1/2}), \quad (6.36)$$

$$\widehat{\Sigma}_B = \mathbf{C}_A^{-1/2} \boldsymbol{\Theta} \mathbf{C}_A^{1/2} \boldsymbol{\Gamma}_{\widehat{\Sigma}_B} \mathbf{C}_A^{1/2} \boldsymbol{\Theta}^\top \mathbf{C}_A^{-1/2} (= \mathbf{C}_B^{1/2} \boldsymbol{\Theta} \mathbf{C}_B^{-1/2} \boldsymbol{\Gamma}_{\widehat{\Sigma}_B} \mathbf{C}_B^{-1/2} \boldsymbol{\Theta}^\top \mathbf{C}_B^{1/2}). \quad (6.37)$$

Here, $\boldsymbol{\Gamma}_{\widehat{\Sigma}_A}$ and $\boldsymbol{\Gamma}_{\widehat{\Sigma}_B}$ are positive-definite diagonal matrices, and $\boldsymbol{\Theta}$ is a block diagonal matrix such that the blocks correspond to the groups sharing the same $c_{a_h} c_{b_h}$, and each block consists of an orthogonal matrix. Furthermore, if there exists a solution with $(\widehat{\Sigma}_A, \widehat{\Sigma}_B)$ written in the form of Eqs. (6.36) and (6.37) with a certain set of $(\boldsymbol{\Gamma}_{\widehat{\Sigma}_A}, \boldsymbol{\Gamma}_{\widehat{\Sigma}_B}, \boldsymbol{\Theta})$, then there also exist its *equivalent* solutions with the same $(\boldsymbol{\Gamma}_{\widehat{\Sigma}_A}, \boldsymbol{\Gamma}_{\widehat{\Sigma}_B})$ for any $\boldsymbol{\Theta}$. Focusing on the solution with $\boldsymbol{\Theta} = \mathbf{I}_H$ as the representative of each *equivalent* class, we can assume that $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$ are diagonal without loss of generality.

6.5 Problem Decomposition

As discussed in Section 6.3, Theorem 6.4 allows us to focus on the solutions that have diagonal posterior covariances, i.e., $\widehat{\Sigma}_A, \widehat{\Sigma}_B \in \mathbb{D}_{++}^H$. For any solution with diagonal covariances, the stationary conditions (6.10) and (6.12) (with Lemma 6.1) imply that $\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}}$ are also diagonal, which means that the column vectors of $\widehat{\mathbf{A}}$, as well as $\widehat{\mathbf{B}}$, are orthogonal to each other. In such a case, the free energy (6.8) depends on the column vectors of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ only through the second term

$$\sigma^{-2} \left\| \mathbf{V} - \widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top \right\|_{\text{Fro}}^2,$$

which coincides with the objective function for the singular value decomposition (SVD). This leads to the following lemma:

Lemma 6.5 *Let*

$$\mathbf{V} = \sum_{h=1}^L \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top \quad (6.38)$$

be the SVD of \mathbf{V} , where γ_h (≥ 0) is the h th largest singular value, and $\boldsymbol{\omega}_{a_h}$ and $\boldsymbol{\omega}_{b_h}$ are the associated right and left singular vectors. Then, any VB solution

(with diagonal posterior covariances) can be written as

$$\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top = \sum_{h=1}^H \widehat{\gamma}_h^{\text{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top \quad (6.39)$$

for some $\{\widehat{\gamma}_h^{\text{VB}} \geq 0\}$.

Thanks to Theorem 6.4 and Lemma 6.5, the variational parameters $\widehat{\mathbf{A}} = (\widehat{\mathbf{a}}_1, \dots, \widehat{\mathbf{a}}_H)$, $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \dots, \widehat{\mathbf{b}}_H)$, $\widehat{\Sigma}_A$, $\widehat{\Sigma}_B$ can be expressed as

$$\begin{aligned}\widehat{\mathbf{a}}_h &= \widehat{a}_h \boldsymbol{\omega}_{a_h}, \\ \widehat{\mathbf{b}}_h &= \widehat{b}_h \boldsymbol{\omega}_{b_h}, \\ \widehat{\Sigma}_A &= \mathbf{Diag}(\widehat{\sigma}_{a_1}^2, \dots, \widehat{\sigma}_{a_H}^2), \\ \widehat{\Sigma}_B &= \mathbf{Diag}(\widehat{\sigma}_{b_1}^2, \dots, \widehat{\sigma}_{b_H}^2),\end{aligned}$$

with a new set of unknowns $\{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H$. Thus, the following holds:

Corollary 6.6 *The VB posterior can be written as*

$$r(\mathbf{A}, \mathbf{B}) = \prod_{h=1}^H \text{Gauss}_M(\mathbf{a}_h; \widehat{a}_h \boldsymbol{\omega}_{a_h}, \widehat{\sigma}_{a_h}^2 \mathbf{I}_M) \prod_{h=1}^H \text{Gauss}_L(\mathbf{b}_h; \widehat{b}_h \boldsymbol{\omega}_{b_h}, \widehat{\sigma}_{b_h}^2 \mathbf{I}_L), \quad (6.40)$$

where $\{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2\}_{h=1}^H$ are the solution of the following minimization problem:

$$\begin{aligned}\text{Given } \sigma^2 \in \mathbb{R}_{++}, \quad \{c_{a_h}^2, c_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H, \\ \min_{\{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2\}_{h=1}^H} F \\ \text{s.t. } \{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H.\end{aligned} \quad (6.41)$$

Here, F is the free energy (6.8), which can be written as

$$2F = LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^H \gamma_h^2}{\sigma^2} + \sum_{h=1}^H 2F_h, \quad (6.42)$$

$$\begin{aligned}\text{where } 2F_h &= M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2}{c_{b_h}^2} \\ &\quad - (L+M) + \frac{-2\widehat{a}_h \widehat{b}_h \gamma_h + (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2)(\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2)}{\sigma^2}.\end{aligned} \quad (6.43)$$

Importantly, the free energy (6.42) depends on the variational parameters $\{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2\}_{h=1}^H$ only through the third term, and the third term is decomposed into H terms, each of which only depends on the variational parameters $(\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2)$ for the h th singular component. Accordingly, given the noise variance σ^2 , we can separately minimize the free energy (6.43), which involves only four unknowns, for each singular component.

6.6 Analytic Form of Global VB Solution

The stationary conditions of Eq. (6.43) are given by

$$\widehat{a}_h = \frac{\widehat{\sigma}_{a_h}^2}{\sigma^2} \gamma_h \widehat{b}_h, \quad (6.44)$$

$$\widehat{\sigma}_{a_h}^2 = \sigma^2 \left(\widehat{b}_h^2 + L \widehat{\sigma}_{b_h}^2 + \frac{\sigma^2}{c_{a_h}^2} \right)^{-1}, \quad (6.45)$$

$$\widehat{b}_h = \frac{\widehat{\sigma}_{b_h}^2}{\sigma^2} \gamma_h \widehat{a}_h, \quad (6.46)$$

$$\widehat{\sigma}_{b_h}^2 = \sigma^2 \left(\widehat{a}_h^2 + M \widehat{\sigma}_{a_h}^2 + \frac{\sigma^2}{c_{b_h}^2} \right)^{-1}, \quad (6.47)$$

which form is a *polynomial system*, a set of polynomial equations, on the four unknowns $(\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2)$. Since Lemma 6.1 guarantees that any minimizer is a stationary point, we can obtain the global solution by finding all points that satisfy the stationary conditions (6.44) through (6.47) and comparing the free energy (6.43) at those points.

This leads to the following theorem and corollary:

Theorem 6.7 *The VB solution is given by*

$$\widehat{\mathbf{U}}^{\text{VB}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \quad \text{where} \quad \widehat{\gamma}_h^{\text{VB}} = \begin{cases} \check{\gamma}_h^{\text{VB}} & \text{if } \gamma_h \geq \check{\gamma}_h^{\text{VB}}, \\ 0 & \text{otherwise,} \end{cases} \quad (6.48)$$

for

$$\check{\gamma}_h^{\text{VB}} = \sigma \sqrt{\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2} + \sqrt{\left(\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2} \right)^2 - LM}}, \quad (6.49)$$

$$\check{\gamma}_h^{\text{VB}} = \gamma_h \left(1 - \frac{\sigma^2}{2\check{\gamma}_h^{\text{VB}}} \left(M + L + \sqrt{(M-L)^2 + \frac{4\check{\gamma}_h^{\text{VB}}}{c_{a_h}^2 c_{b_h}^2}} \right) \right). \quad (6.50)$$

Corollary 6.8 *The VB posterior is given by Eq. (6.40) with the following variational parameters: if $\gamma_h > \underline{\gamma}_h^{\text{VB}}$,*

$$\widehat{a}_h = \pm \sqrt{\check{\gamma}_h^{\text{VB}} \widehat{\delta}_h^{\text{VB}}}, \quad \widehat{b}_h = \pm \sqrt{\frac{\check{\gamma}_h^{\text{VB}}}{\widehat{\delta}_h^{\text{VB}}}}, \quad \widehat{\sigma}_{a_h}^2 = \frac{\sigma^2 \widehat{\delta}_h^{\text{VB}}}{\gamma_h}, \quad \widehat{\sigma}_{b_h}^2 = \frac{\sigma^2}{\gamma_h \widehat{\delta}_h^{\text{VB}}}, \quad (6.51)$$

$$\text{where } \widehat{\delta}_h^{\text{VB}} \left(\equiv \frac{\widehat{a}_h}{\widehat{b}_h} \right) = \frac{c_{a_h}}{\sigma^2} \left(\gamma_h - \check{\gamma}_h^{\text{VB}} - \frac{L\sigma^2}{\gamma_h} \right), \quad (6.52)$$

and otherwise,

$$\widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = c_{a_h}^2 \left(1 - \frac{L\widehat{\zeta}_h^{\text{VB}}}{\sigma^2} \right), \quad \widehat{\sigma}_{b_h}^2 = c_{b_h}^2 \left(1 - \frac{M\widehat{\zeta}_h^{\text{VB}}}{\sigma^2} \right), \quad (6.53)$$

where

$$\widehat{\zeta}_h^{\text{VB}} \left(\equiv \widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2 \right) = \frac{\sigma^2}{2LM} \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} - \sqrt{\left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right)^2 - 4LM} \right). \quad (6.54)$$

Theorem 6.7 states that the VB solution for fully observed MF is a truncated shrinkage SVD with the truncation threshold and the shrinkage estimator given by Eqs. (6.49) and (6.50), respectively. Corollary 6.8 completely specifies the VB posterior.¹

These results give insights into the behavior of VB learning; for example, they explain why a sparse solution is obtained, and what are similarities and differences between the Bayes posterior and the VB posterior, which will be discussed in Chapter 7. The results also form the basis of further analysis on the global empirical VB solution (Section 6.8), which will be used for performance guarantee (Chapter 8), and global (or efficient local) solvers for multilinear models (Chapters 9, 10, and 11). Before moving on, we give the proofs of the theorem and the corollary in the next section.

6.7 Proofs of Theorem 6.7 and Corollary 6.8

We will find all stationary points that satisfy Eqs. (6.44) through (6.47), and compare the free energy (6.43).

¹ The similarity between $(\underline{\gamma}_h^{\text{VB}})^2$ and $LM\widehat{\zeta}_h^{\text{VB}}$ comes from the fact that they are the two different solutions of the same quadratic equations, i.e., Eq. (6.79) with respect to $(\underline{\gamma}_h^{\text{VB}})^2$ and (6.77) with respect to $LM\widehat{\zeta}_h^{\text{VB}}$.

By using Eqs. (6.45) and (6.47), the free energy (6.43) can be simplified as

$$\begin{aligned} F_h &= M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{1}{\sigma^2} \left(a_h^2 + M \widehat{\sigma}_{a_h}^2 + \frac{\sigma^2}{c_{b_h}^2} \right) \left(b_h^2 + L \widehat{\sigma}_{b_h}^2 + \frac{\sigma^2}{c_{a_h}^2} \right) \\ &\quad - (L + M) + \frac{-2a_h b_h \gamma_h}{\sigma^2} - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \\ &= M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{\sigma^2}{\widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2} - \frac{2\widehat{a}_h \widehat{b}_h \gamma_h}{\sigma^2} - \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right). \end{aligned} \quad (6.55)$$

The stationary conditions (6.44) through (6.47) imply two possibilities of stationary points.

6.7.1 Null Stationary Point

If $\widehat{a}_h = 0$ or $\widehat{b}_h = 0$, Eqs. (6.44) and (6.46) require that $\widehat{a}_h = 0$ and $\widehat{b}_h = 0$. In this case, Eqs. (6.45) and (6.47) lead to

$$\widehat{\sigma}_{a_h}^2 = c_{a_h}^2 \left(1 - \frac{L \widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2}{\sigma^2} \right), \quad (6.56)$$

$$\widehat{\sigma}_{b_h}^2 = c_{b_h}^2 \left(1 - \frac{M \widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2}{\sigma^2} \right). \quad (6.57)$$

Multiplying Eqs. (6.56) and (6.57), we have

$$\left(1 - \frac{L \widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2}{\sigma^2} \right) \left(1 - \frac{M \widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2}{\sigma^2} \right) = \frac{\widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2}{c_{a_h}^2 c_{b_h}^2}, \quad (6.58)$$

and therefore

$$\frac{LM}{\sigma^2} \sigma_{a_h}^4 \sigma_{b_h}^4 - \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right) \widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2 + \sigma^2 = 0. \quad (6.59)$$

Solving the quadratic equation (6.59) with respect to $\widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2$ and checking the signs of $\widehat{\sigma}_{a_h}^2$ and $\widehat{\sigma}_{b_h}^2$, we have the following lemma:

Lemma 6.9 *For any $\gamma_h \geq 0$ and $c_{a_h}^2, c_{b_h}^2, \sigma^2 \in \mathbb{R}_{++}$, the null stationary point given by Eq. (6.53) exists with the following free energy:*

$$F_h^{\text{VB-Null}} = -M \log \left(1 - \frac{L}{\sigma^2} \widehat{\zeta}_h^{\text{VB}} \right) - L \log \left(1 - \frac{M}{\sigma^2} \widehat{\zeta}_h^{\text{VB}} \right) - \frac{LM}{\sigma^2} \widehat{\zeta}_h^{\text{VB}}, \quad (6.60)$$

where $\widehat{\zeta}_h^{\text{VB}}$ is defined by Eq. (6.54).

Proof Eq. (6.59) has two positive real solutions:

$$\widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2 = \frac{\sigma^2}{2LM} \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \pm \sqrt{\left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right)^2 - 4LM} \right).$$

The larger solution (with the plus sign) is decreasing with respect to $c_{a_h}^2 c_{b_h}^2$, and lower-bounded as $\widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2 > \sigma^2/L$. The smaller solution (with the minus sign) is increasing with respect to $c_{a_h}^2 c_{b_h}^2$, and upper-bounded as $\widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2 < \sigma^2/M$.

For $\widehat{\sigma}_{a_h}^2$ and $\widehat{\sigma}_{b_h}^2$ to be positive, Eqs. (6.56) and (6.57) require that

$$\widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2 < \frac{\sigma^2}{M}, \quad (6.61)$$

which is violated by the larger solution, while satisfied by the smaller solution. With the smaller solution (6.54), Eqs. (6.56) and (6.57) give the stationary point given by Eq. (6.53).

Using Eq. (6.59), we can easily derive Eq. (6.60) from Eq. (6.55), which completes the proof of Lemma 6.9.

□

6.7.2 Positive Stationary Point

Assume that $\widehat{a}_h, \widehat{b}_h \neq 0$. In this case, Eqs. (6.44) and (6.46) imply that \widehat{a}_h and \widehat{b}_h have the same sign. Define

$$\widehat{\gamma}_h = \widehat{a}_h \widehat{b}_h > 0, \quad (6.62)$$

$$\widehat{\delta}_h = \frac{\widehat{a}_h}{\widehat{b}_h} > 0. \quad (6.63)$$

From Eqs. (6.44) and (6.46), we have

$$\sigma_{a_h}^2 = \frac{\sigma^2}{\gamma_h} \widehat{\delta}_h, \quad (6.64)$$

$$\sigma_{b_h}^2 = \frac{\sigma^2}{\gamma_h} \widehat{\delta}_h^{-1}. \quad (6.65)$$

Substituting Eqs. (6.64) and (6.65) into Eqs. (6.45) and (6.47) gives

$$\gamma_h \widehat{\delta}_h^{-1} = \left(\widehat{\gamma}_h \widehat{\delta}_h^{-1} + L \frac{\sigma^2}{\gamma_h} \widehat{\delta}_h^{-1} + \frac{\sigma^2}{c_{a_h}^2} \right), \quad (6.66)$$

$$\gamma_h \widehat{\delta}_h = \left(\widehat{\gamma}_h \widehat{\delta}_h + M \frac{\sigma^2}{\gamma_h} \widehat{\delta}_h + \frac{\sigma^2}{c_{b_h}^2} \right), \quad (6.67)$$

and therefore,

$$\widehat{\delta}_h = \frac{c_{a_h}}{\sigma^2} \left(\gamma_h - \widehat{\gamma}_h - \frac{L\sigma^2}{\gamma_h} \right), \quad (6.68)$$

$$\widehat{\delta}_h^{-1} = \frac{c_{b_h}}{\sigma^2} \left(\gamma_h - \widehat{\gamma}_h - \frac{M\sigma^2}{\gamma_h} \right). \quad (6.69)$$

Multiplying Eqs. (6.68) and (6.69), we have

$$\left(\gamma_h - \widehat{\gamma}_h - \frac{L\sigma^2}{\gamma_h} \right) \left(\gamma_h - \widehat{\gamma}_h - \frac{M\sigma^2}{\gamma_h} \right) = \frac{\sigma^4}{c_{a_h} c_{b_h}}, \quad (6.70)$$

and therefore

$$\widehat{\gamma}_h^2 - \left(2\gamma_h - \frac{(L+M)\sigma^2}{\gamma_h} \right) \widehat{\gamma}_h + \left(\gamma_h - \frac{L\sigma^2}{\gamma_h} \right) \left(\gamma_h - \frac{M\sigma^2}{\gamma_h} \right) - \frac{\sigma^4}{c_{a_h} c_{b_h}} = 0. \quad (6.71)$$

By solving the quadratic equation (6.71) with respect to $\widehat{\gamma}_h$, and checking the signs of $\widehat{\gamma}_h$, $\widehat{\delta}_h$, $\widehat{\sigma}_{a_h}^2$, and $\widehat{\sigma}_{b_h}^2$, we have the following lemma:

Lemma 6.10 *If and only if $\gamma_h > \underline{\gamma}_h^{\text{VB}}$, where $\underline{\gamma}_h^{\text{VB}}$ is defined by Eq. (6.49), the positive stationary point given by Eq. (6.51) exists with the following free energy:*

$$\begin{aligned} F_h^{\text{VB-Posi}} = & -M \log \left(1 - \left(\frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{L\sigma^2}{\gamma_h^2} \right) \right) - L \log \left(1 - \left(\frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{M\sigma^2}{\gamma_h^2} \right) \right) \\ & - \frac{\gamma_h^2}{\sigma^2} \left(\frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{L\sigma^2}{\gamma_h^2} \right) \left(\frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{M\sigma^2}{\gamma_h^2} \right), \end{aligned} \quad (6.72)$$

where $\check{\gamma}_h^{\text{VB}}$ is defined by Eq. (6.50).

Proof Since $\widehat{\delta}_h > 0$, Eqs. (6.68) and (6.69) require that

$$\widehat{\gamma}_h < \gamma_h - \frac{L\sigma^2}{\gamma_h}, \quad (6.73)$$

and therefore, the positive stationary point exists only when

$$\gamma_h > \sqrt{M}\sigma. \quad (6.74)$$

Let us assume that Eq. (6.74) holds.

Eq. (6.71) has two solutions:

$$\widehat{\gamma}_h = \frac{1}{2} \left(2\gamma_h - \frac{(L+M)\sigma^2}{\gamma_h} \pm \sqrt{\left(\frac{(M-L)\sigma^2}{\gamma_h} \right)^2 + \frac{4\sigma^4}{c_{a_h}^2 c_{b_h}^2}} \right).$$

The larger solution with the plus sign is positive, decreasing with respect to $c_{a_h}^2 c_{b_h}^2$, and lower-bounded as $\widehat{\gamma}_h > \gamma_h - L\sigma^2/\gamma_h$, which violates the condition (6.73).

The smaller solution, Eq. (6.50), with the minus sign is positive if the intercept of the left-hand side in Eq. (6.71) is positive, i.e.,

$$\left(\gamma_h - \frac{L\sigma^2}{\gamma_h}\right)\left(\gamma_h - \frac{M\sigma^2}{\gamma_h}\right) - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} > 0. \quad (6.75)$$

From the condition (6.75), we obtain the threshold (6.49) for the existence of the positive stationary point. Note that $\underline{\gamma}_h^{\text{VB}} > \sqrt{M}\sigma$, and therefore, Eq. (6.74) holds whenever $\gamma_h > \underline{\gamma}_h^{\text{VB}}$.

Assume that $\gamma_h > \underline{\gamma}_h^{\text{VB}}$. Then, with the solution (6.50), $\widehat{\delta}_h$, given by Eq. (6.68), and $\widehat{\sigma}_{a_h}^2$ and $\widehat{\sigma}_{b_h}^2$, given by Eqs. (6.64) and (6.65), are all positive. Thus, we obtain the positive stationary point (6.51).

Substituting Eqs. (6.64) and (6.65), and then Eqs. (6.68) and (6.69), into the free energy (6.55), we have

$$\begin{aligned} F_h^{\text{VB-Posi}} = & -M \log \left(1 - \frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} - \frac{L\sigma^2}{\gamma_h^2} \right) - L \log \left(1 - \frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} - \frac{M\sigma^2}{\gamma_h^2} \right) \\ & + \frac{-2\gamma_h \check{\gamma}_h^{\text{VB}}}{\sigma^2} + \frac{\gamma_h^2}{\sigma^2} - \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right). \end{aligned} \quad (6.76)$$

Using Eq. (6.70), we can eliminate the direct dependency on $c_{a_h}^2 c_{b_h}^2$, and express the free energy (6.76) as a function of $\check{\gamma}_h^{\text{VB}}$. This results in Eq. (6.72), and completes the proof of Lemma 6.10. \square

6.7.3 Useful Relations

Let us summarize some useful relations between variables, which are used in the subsequent sections. $\widehat{\zeta}_h^{\text{VB}}$, $\check{\gamma}_h^{\text{VB}}$, and $\underline{\gamma}_h^{\text{VB}}$, derived from Eqs. (6.58), (6.70), and the constant part of Eq. (6.71), respectively, satisfy the following:

$$\left(1 - \frac{L\widehat{\zeta}_h^{\text{VB}}}{\sigma^2}\right)\left(1 - \frac{M\widehat{\zeta}_h^{\text{VB}}}{\sigma^2}\right) - \frac{\widehat{\zeta}_h^{\text{VB}}}{c_{a_h}^2 c_{b_h}^2} = 0, \quad (6.77)$$

$$\left(\gamma_h - \check{\gamma}_h^{\text{VB}} - \frac{L\sigma^2}{\gamma_h}\right)\left(\gamma_h - \check{\gamma}_h^{\text{VB}} - \frac{M\sigma^2}{\gamma_h}\right) - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} = 0, \quad (6.78)$$

$$\left(\underline{\gamma}_h^{\text{VB}} - \frac{L\sigma^2}{\underline{\gamma}_h^{\text{VB}}}\right)\left(\underline{\gamma}_h^{\text{VB}} - \frac{M\sigma^2}{\underline{\gamma}_h^{\text{VB}}}\right) - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} = 0. \quad (6.79)$$

From Eqs. (6.54) and (6.49), we find that

$$\underline{\gamma}_h^{\text{VB}} = \sqrt{\left((L + M)\sigma^2 + \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} \right) - LM\widehat{\zeta}_h^{\text{VB}}}, \quad (6.80)$$

which is useful when comparing the free energies of the null and the positive stationary points.

6.7.4 Free Energy Comparison

Lemmas 6.9 and 6.10 imply that, when $\gamma_h \leq \underline{\gamma}_h^{\text{VB}}$, the null stationary point is only the stationary point, and therefore the global solution. When $\gamma_h > \underline{\gamma}_h^{\text{VB}}$, both of the null and the positive stationary points exist, and therefore identifying the global solution requires us to compare their free energies, given by Eqs. (6.60) and (6.72).

Given the observed singular value $\gamma_h \geq 0$, we view the free energy as a function of $c_{a_h}^2 c_{b_h}^2$. We also view the threshold $\underline{\gamma}_h^{\text{VB}}$ as a function of $c_{a_h}^2 c_{b_h}^2$. We find from Eq. (6.49) that $\underline{\gamma}_h^{\text{VB}}$ is decreasing and lower-bounded by $\underline{\gamma}_h^{\text{VB}} > \sqrt{M}\sigma$. Therefore, when $\gamma_h \leq \sqrt{M}\sigma$, $\underline{\gamma}_h^{\text{VB}}$ never gets smaller than γ_h for any $c_{a_h}^2 c_{b_h}^2 > 0$. When $\gamma_h > \sqrt{M}\sigma$, on the other hand, there is a threshold $\underline{c}_{a_h} \underline{c}_{b_h}$ such that $\gamma_h > \underline{\gamma}_h^{\text{VB}}$ if $c_{a_h}^2 c_{b_h}^2 > \underline{c}_{a_h} \underline{c}_{b_h}$. Eq. (6.79) implies that the threshold is given by

$$\underline{c}_{a_h} \underline{c}_{b_h} = \frac{\sigma^4}{\gamma_h^2 \left(1 - \frac{L\sigma^2}{\gamma_h^2} \right) \left(1 - \frac{M\sigma^2}{\gamma_h^2} \right)}. \quad (6.81)$$

We have the following lemma:

Lemma 6.11 *For any $\gamma_h \geq 0$ and $c_{a_h}^2 c_{b_h}^2 > 0$, the derivative of the free energy (6.60) at the null stationary point with respect to $c_{a_h}^2 c_{b_h}^2$ is given by*

$$\frac{\partial F_h^{\text{VB-Null}}}{\partial c_{a_h}^2 c_{b_h}^2} = \frac{LM\widehat{\zeta}_h^{\text{VB}}}{\sigma^2 c_{a_h}^2 c_{b_h}^2}. \quad (6.82)$$

For $\gamma_h > M/\sigma^2$ and $c_{a_h}^2 c_{b_h}^2 > \underline{c}_{a_h} \underline{c}_{b_h}$, the derivative of the free energy (6.72) at the positive stationary point with respect to $c_{a_h}^2 c_{b_h}^2$ is given by

$$\frac{\partial F_h^{\text{VB-Posi}}}{\partial c_{a_h}^2 c_{b_h}^2} = \frac{\gamma_h^2}{\sigma^2 c_{a_h}^2 c_{b_h}^2} \left(\frac{(\check{\gamma}_h^{\text{VB}})^2}{\gamma_h^2} - \left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right) \frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{LM\sigma^4}{\gamma_h^4} \right). \quad (6.83)$$

The derivative of the difference is negative, i.e.,

$$\frac{\partial(F_h^{\text{VB-Posi}} - F_h^{\text{VB-Null}})}{\partial c_{a_h}^2 c_{b_h}^2} = -\frac{1}{\sigma^2 c_{a_h}^2 c_{b_h}^2} \left(\gamma_h (\gamma_h - \check{\gamma}_h^{\text{VB}}) - (\underline{\gamma}_h^{\text{VB}})^2 \right) < 0. \quad (6.84)$$

Proof By differentiating Eqs. (6.60), (6.54), (6.72), and (6.50), we have

$$\begin{aligned} \frac{\partial F_h^{\text{VB-Null}}}{\partial \check{\zeta}_h^{\text{VB}}} &= \frac{LM}{\sigma^2 \left(1 - \frac{L}{\sigma^2} \check{\zeta}_h^{\text{VB}}\right)} + \frac{LM}{\sigma^2 \left(1 - \frac{M}{\sigma^2} \check{\zeta}_h^{\text{VB}}\right)} - \frac{LM}{\sigma^2} \\ &= \frac{LM c_{a_h}^2 c_{b_h}^2 \left(1 + \frac{\sqrt{LM}}{\sigma^2} \check{\zeta}_h^{\text{VB}}\right) \left(1 - \frac{\sqrt{LM}}{\sigma^2} \check{\zeta}_h^{\text{VB}}\right)}{\sigma^2 \check{\zeta}_h^{\text{VB}}}, \end{aligned} \quad (6.85)$$

$$\begin{aligned} \frac{\partial \check{\zeta}_h^{\text{VB}}}{\partial c_{a_h}^2 c_{b_h}^2} &= \frac{\sigma^2}{2LM} \left(-\frac{\sigma^2}{c_{a_h}^4 c_{b_h}^4} + \frac{2\sigma^2 \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}\right)}{2c_{a_h}^4 c_{b_h}^4 \sqrt{\left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}\right)^2 - 4LM}} \right) \\ &= \frac{1}{c_{a_h}^4 c_{b_h}^4} \left(\frac{(\check{\zeta}_h^{\text{VB}})^2}{\left(1 - \frac{\sqrt{LM}\check{\zeta}_h^{\text{VB}}}{\sigma^2}\right) \left(1 + \frac{\sqrt{LM}\check{\zeta}_h^{\text{VB}}}{\sigma^2}\right)} \right), \end{aligned} \quad (6.86)$$

$$\begin{aligned} \frac{\partial F_h^{\text{VB-Posi}}}{\partial \check{\gamma}_h^{\text{VB}}} &= \frac{M}{\gamma_h \left(1 - \left(\frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{L\sigma^2}{\gamma_h^2}\right)\right)} + \frac{L}{\gamma_h \left(1 - \left(\frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{M\sigma^2}{\gamma_h^2}\right)\right)} \\ &\quad - \frac{\gamma_h}{\sigma^2} \left(\frac{2\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{(L+M)\sigma^2}{\gamma_h^2} \right) \\ &= \frac{2c_{a_h}^2 c_{b_h}^2 \gamma_h^3 \left(1 - \left(\frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{(L+M)\sigma^2}{2\gamma_h^2}\right)\right) \left(\frac{(\check{\gamma}_h^{\text{VB}})^2}{\gamma_h^2} - \left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right) \frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{LM\sigma^4}{\gamma_h^4}\right)}{\sigma^6}, \end{aligned} \quad (6.87)$$

$$\begin{aligned} \frac{\partial \check{\gamma}_h}{\partial c_{a_h}^2 c_{b_h}^2} &= \frac{4\gamma_h^2 \sigma^2}{4\gamma_h c_{a_h}^4 c_{b_h}^4 \sqrt{(M-L)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}}} \\ &= \frac{\sigma^4}{2\gamma_h c_{a_h}^4 c_{b_h}^4 \left(1 - \left(\frac{\check{\gamma}_h^{\text{VB}}}{\gamma_h} + \frac{(L+M)\sigma^2}{2\gamma_h^2}\right)\right)}. \end{aligned} \quad (6.88)$$

Here, we used Eqs. (6.54) and (6.77) to obtain Eqs. (6.85) and (6.86), and Eqs. (6.50) and (6.78) to obtain Eqs. (6.87) and (6.88), respectively. Eq. (6.82) is obtained by multiplying Eqs. (6.85) and (6.86), while Eq. (6.83) is obtained by multiplying Eqs. (6.87) and (6.88).

Taking the difference between the derivatives (6.82) and (6.83), and then using Eqs. (6.78) and (6.80), we have

$$\begin{aligned} \frac{\partial(F_h^{\text{VB-Posi}} - F_h^{\text{VB-Null}})}{\partial c_{a_h}^2 c_{b_h}^2} &= \frac{\partial F_h^{\text{VB-Posi}}}{\partial c_{a_h}^2 c_{b_h}^2} - \frac{\partial F_h^{\text{VB-Null}}}{\partial c_{a_h}^2 c_{b_h}^2} \\ &= -\frac{1}{\sigma^2 c_{a_h}^2 c_{b_h}^2} (\gamma_h (\gamma_h - \check{\gamma}_h) - (\underline{\gamma}_h^{\text{VB}})^2). \end{aligned} \quad (6.89)$$

The following can be obtained from Eqs. (6.78) and (6.79), respectively:

$$\left(\gamma_h (\gamma_h - \check{\gamma}_h^{\text{VB}}) - \frac{(L+M)\sigma^2}{2} \right)^2 = \frac{(L+M)^2 \sigma^4}{4} - LM\sigma^4 + \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} \gamma_h^2, \quad (6.90)$$

$$\left((\underline{\gamma}_h^{\text{VB}})^2 - \frac{(L+M)\sigma^2}{2} \right)^2 = \frac{(L+M)^2 \sigma^4}{4} - LM\sigma^4 + \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} (\underline{\gamma}_h^{\text{VB}})^2. \quad (6.91)$$

Eqs. (6.90) and (6.91) imply that

$$\gamma_h (\gamma_h - \check{\gamma}_h^{\text{VB}}) > (\underline{\gamma}_h^{\text{VB}})^2 \quad \text{when} \quad \gamma_h > \underline{\gamma}_h^{\text{VB}}.$$

Therefore, Eq. (6.89) is negative, which completes the proof of Lemma 6.11. \square

It is easy to show that the null stationary point (6.53) and the positive stationary point (6.51) coincide with each other at $c_{a_h}^2 c_{b_h}^2 \rightarrow \underline{c}_{a_h} \underline{c}_{b_h} + 0$. Here, $+0$ means that it approaches to zero from the positive side. Therefore,

$$\lim_{c_{a_h}^2 c_{b_h}^2 \rightarrow \underline{c}_{a_h} \underline{c}_{b_h} + 0} (F_h^{\text{VB-Posi}} - F_h^{\text{VB-Null}}) = 0. \quad (6.92)$$

Eqs. (6.84) and (6.92) together imply that

$$F_h^{\text{VB-Posi}} - F_h^{\text{VB-Null}} < 0 \quad \text{for} \quad c_{a_h}^2 c_{b_h}^2 > \underline{c}_{a_h} \underline{c}_{b_h}, \quad (6.93)$$

which results in the following lemma:

Lemma 6.12 *The positive stationary point is the global solution (the global minimizer of the free energy (6.43) for fixed c_{a_h} and c_{b_h}) whenever it exists.*

Combining Lemmas 6.9, 6.10, and 6.12 completes the proof of Theorem 6.7 and Corollary 6.8. \square

Figure 6.3 illustrates the behavior of the free energies.

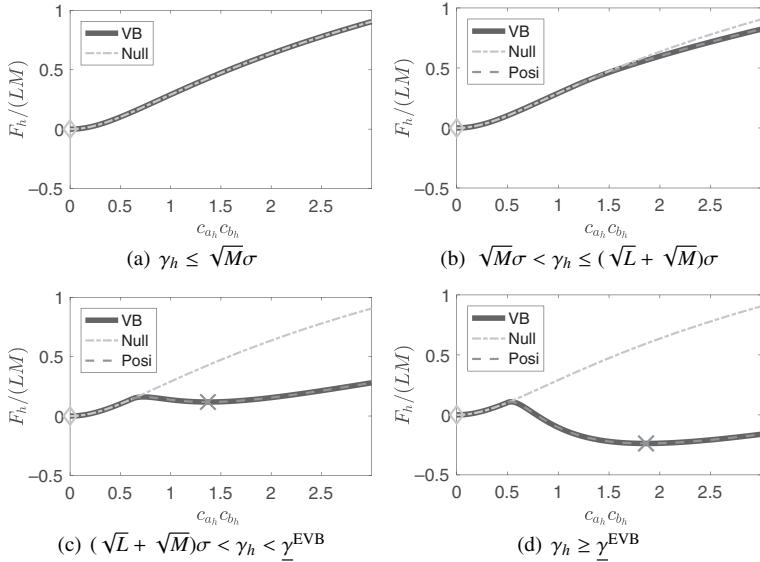


Figure 6.3 Behavior of the free energies (6.60) and (6.72) at the null and the positive stationary points as functions of $c_{a_h}c_{b_h}$, when $L = M = H = 1$ and $\sigma^2 = 1$. The curve labeled as ‘‘VB’’ shows the VB free energy $F_h = \min(F_h^{\text{VB-Null}}, F_h^{\text{VB-Posi}})$ at the global solution, given $c_{a_h}c_{b_h}$. If $\gamma_h \leq \sqrt{M}\sigma$, only the null stationary point exists for any $c_{a_h}c_{b_h} > 0$. Otherwise, the positive stationary point exists for $c_{a_h}c_{b_h} > \underline{c}_{a_h}\underline{c}_{b_h}$, and it is the global minimum whenever it exists. In empirical VB learning where $c_{a_h}c_{b_h}$ is also optimized, $c_{a_h}c_{b_h} \rightarrow 0$ (indicated by a diamond) is the unique local minimum if $\gamma_h \leq (\sqrt{L} + \sqrt{M})\sigma$. Otherwise, a positive local minimum also exists (indicated by a cross), and it is the global minimum if and only if $\gamma_h \geq \underline{\gamma}^{\text{EVB}}$ (see Section 6.9 for detailed discussion).

6.8 Analytic Form of Global Empirical VB Solution

In this section, we will solve the empirical VB (EVB) problem where the prior covariances are also estimated from observation, i.e.,

$$\widehat{\boldsymbol{r}} = \underset{\boldsymbol{r}, \mathbf{C}_A, \mathbf{C}_B}{\operatorname{argmin}} F(\boldsymbol{r}) \quad \text{s.t.} \quad \boldsymbol{r}(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}). \quad (6.94)$$

Since the solution of the EVB problem is a VB solution with some values for the prior covariances $\mathbf{C}_A, \mathbf{C}_B$, the empirical VB posterior is in the same form as the VB posterior (6.40). Accordingly, the problem (6.94) can be written with the variational parameters $\{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2\}_{h=1}^H$ as follows:

Given $\sigma^2 \in \mathbb{R}_{++}$,

$$\min_{\{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2, c_{a_h}^2, c_{b_h}^2\}_{h=1}^H} F \quad (6.95)$$

$$\text{s.t. } \{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2, c_{a_h}^2, c_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H, \quad (6.96)$$

where the free energy F is given by Eq. (6.42).

Solving the empirical VB problem (6.95) is not much harder than the VB problem (6.41) because the objective is still separable into H singular components when the prior variances $\{c_{a_h}^2, c_{b_h}^2\}$ are also optimized. More specifically, we can obtain the empirical VB solution by minimizing the *componentwise free energy* (6.43) with respect to the only six unknowns $(\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2, c_{a_h}^2, c_{b_h}^2)$ for each h th component. On the other hand, analyzing the VB estimator for the noise variance σ^2 is hard, since F_h for all $h = 1, \dots, H$ depend on σ^2 and therefore the free energy (6.42) is not separable. We postpone the analysis of this full empirical VB learning to Chapter 8, where the theoretical performance guarantee is derived.

For the problem (6.95), the stationary points of the free energy (6.43) satisfy Eqs. (6.44) through (6.47) along with Eqs. (3.26) and (3.27), which are written with the new set of variational parameters as

$$c_{a_h}^2 = \widehat{a}_h^2/M + \widehat{\sigma}_{a_h}^2, \quad (6.97)$$

$$c_{b_h}^2 = \widehat{b}_h^2/L + \widehat{\sigma}_{b_h}^2. \quad (6.98)$$

However, unlike the VB solution, for which Lemma 6.1 holds, we cannot assume that the EVB solution is a stationary point, since the free energy F_h given by Eq. (6.43) does not necessarily diverge to $+\infty$ when approaching the domain boundary (6.96). More specifically, F_h can converge to a finite value, for example, for $\widehat{a}_h = \widehat{b}_h = 0, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2, c_{a_h}^2, c_{b_h}^2 \rightarrow +0$. Taking this into account, we can obtain the following theorem:

Theorem 6.13 *Let*

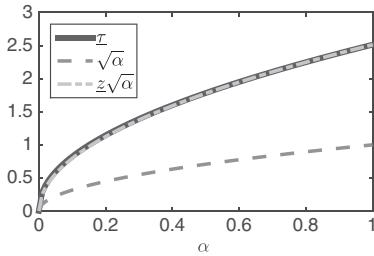
$$\alpha = \frac{L}{M} \quad (0 < \alpha \leq 1), \quad (6.99)$$

and let $\underline{\tau} = \underline{\tau}(\alpha)$ be the unique zero-cross point of the following decreasing function:

$$\Xi(\tau; \alpha) = \Phi(\tau) + \Phi\left(\frac{\tau}{\alpha}\right), \quad \text{where} \quad \Phi(z) = \frac{\log(z+1)}{z} - \frac{1}{2}. \quad (6.100)$$

Then, the EVB solution is given by

$$\widehat{\boldsymbol{U}}^{\text{EVB}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{EVB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \quad \text{where} \quad \widehat{\gamma}_h^{\text{EVB}} = \begin{cases} \check{\gamma}_h^{\text{EVB}} & \text{if } \gamma_h \geq \underline{\gamma}^{\text{EVB}}, \\ 0 & \text{otherwise,} \end{cases} \quad (6.101)$$

Figure 6.4 Values of $\underline{\tau}(\alpha)$, $\sqrt{\alpha}$, and $\underline{z}\sqrt{\alpha}$.

for

$$\underline{\gamma}^{\text{EVB}} = \sigma \sqrt{M \left(1 + \underline{\tau}\right) \left(1 + \frac{\alpha}{\underline{\tau}}\right)}, \quad (6.102)$$

$$\check{\gamma}_h^{\text{EVB}} = \frac{\gamma_h}{2} \left(1 - \frac{(M+L)\sigma^2}{\gamma_h^2} + \sqrt{\left(1 - \frac{(M+L)\sigma^2}{\gamma_h^2} \right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}} \right). \quad (6.103)$$

The EVB threshold (6.102) involves $\underline{\tau}$, which needs to be numerically computed. However, we can easily prepare a table of the values for $0 < \alpha \leq 1$ beforehand, like the cumulative Gaussian probability used in statistical tests (Pearson, 1914). Alternatively, $\underline{\tau} \approx \underline{z}\sqrt{\alpha}$ is a good approximation, where $\underline{z} \approx 2.5129$ is the unique zero-cross point of $\Phi(z)$, as seen in Figure 6.4. We can show that $\underline{\tau}$ lies in the following range (Lemma 6.18 in Section 6.9):

$$\sqrt{\alpha} < \underline{\tau} < \underline{z}. \quad (6.104)$$

We will see in Chapter 8 that $\underline{\tau}$ is an important quantity in describing the behavior of the full empirical VB solution where the noise variance σ^2 is also estimated from observation.

In Section 6.9, we give the proof of Theorem 6.13. Then, in Section 6.10, some corollaries obtained and variables defined in the proof are summarized, which will be used in Chapter 8.

6.9 Proof of Theorem 6.13

In this section, we prove Theorem 6.13, which provides explicit forms, Eqs. (6.102) and (6.103), of the EVB threshold $\underline{\gamma}^{\text{EVB}}$ and the EVB shrinkage estimator $\check{\gamma}_h^{\text{EVB}}$. In fact, we can easily obtain Eq. (6.103) in an intuitive way, by

using some of the results obtained in Section 6.7. After that, by expressing the free energy F_h with normalized versions of the observation and the estimator, we derive Eq. (6.102).

6.9.1 EVB Shrinkage Estimator

Eqs. (6.60) and (6.72) imply that the free energy does not depend on the ratio c_{a_h}/c_{b_h} between the hyperparameters. Accordingly, we fix the ratio to $c_{a_h}/c_{b_h} = 1$ without loss of generality. Lemma 6.11 allows us to minimize the free energy with respect to $c_{a_h}c_{b_h}$ in a straightforward way.

Let us regard the free energies (6.60) and (6.72) at the null and the positive stationary points as functions of $c_{a_h}c_{b_h}$ (see Figure 6.3). Then, we find from Eq. (6.82) that

$$\frac{\partial F_h^{\text{VB-Null}}}{\partial c_{a_h}^2 c_{b_h}^2} > 0,$$

which implies that the free energy (6.60) at the null stationary point is increasing. Using Lemma 6.9, we thus have the following lemma:

Lemma 6.14 *For any given $\gamma_h \geq 0$ and $\sigma^2 > 0$, the null EVB local solution, given by*

$$\widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = \sqrt{\zeta^{\text{EVB}}}, \quad \widehat{\sigma}_{b_h}^2 = \sqrt{\zeta^{\text{EVB}}}, \quad c_{a_h}c_{b_h} = \sqrt{\zeta^{\text{EVB}}},$$

where $\zeta^{\text{EVB}} \rightarrow +0,$

exists, and its free energy is given by

$$F_h^{\text{EVB-Null}} \rightarrow +0. \quad (6.105)$$

When $\gamma_h \geq (\sqrt{L} + \sqrt{M})\sigma$, the derivative (6.83) of the free energy (6.72) at the positive stationary point can be further factorized as

$$\frac{\partial F_h^{\text{VB-Posi}}}{\partial c_{a_h}^2 c_{b_h}^2} = \frac{\gamma_h}{\sigma^2 c_{a_h}^2 c_{b_h}^2} (\check{\gamma}_h^{\text{VB}} - \check{\gamma}_h)(\check{\gamma}_h^{\text{VB}} - \check{\gamma}_h^{\text{EVB}}), \quad (6.106)$$

$$\text{where } \check{\gamma}_h = \frac{\gamma_h}{2} \left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2} - \sqrt{\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2} \right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}} \right), \quad (6.107)$$

and $\check{\gamma}_h^{\text{EVB}}$ is given by Eq. (6.103). The VB shrinkage estimator (6.50) is an increasing function of $c_{a_h}c_{b_h}$ ranging over

$$0 < \check{\gamma}_h^{\text{VB}} < \gamma_h - \frac{M\sigma^2}{\gamma_h},$$

and both of Eqs. (6.107) and (6.103) are in this range, i.e.,

$$0 < \check{\gamma}_h \leq \check{\gamma}_h^{\text{EVB}} < \gamma_h - \frac{M\sigma^2}{\gamma_h}.$$

Therefore Eq. (6.106) leads to the following lemma:

Lemma 6.15 *If $\gamma_h \leq (\sqrt{L} + \sqrt{M})\sigma$, the free energy $F_h^{\text{VB-Posi}}$ at the positive stationary point is monotonically increasing. Otherwise,*

$$F_h^{\text{VB-Posi}} \text{ is } \begin{cases} \text{increasing} & \text{for } \check{\gamma}_h^{\text{VB}} < \check{\gamma}_h, \\ \text{decreasing} & \text{for } \check{\gamma}_h < \check{\gamma}_h^{\text{VB}} < \check{\gamma}_h^{\text{EVB}}, \\ \text{increasing} & \text{for } \check{\gamma}_h^{\text{VB}} > \check{\gamma}_h^{\text{EVB}}, \end{cases}$$

and therefore minimized at $\check{\gamma}_h^{\text{VB}} = \check{\gamma}_h^{\text{EVB}}$.

We can see this behavior of the free energy in Figure 6.3.

The derivative (6.83) is zero when $\check{\gamma}_h^{\text{VB}} = \check{\gamma}_h^{\text{EVB}}$, which leads to

$$\left(\check{\gamma}_h^{\text{EVB}} + \frac{L\sigma^2}{\gamma_h}\right)\left(\check{\gamma}_h^{\text{EVB}} + \frac{M\sigma^2}{\gamma_h}\right) = \gamma_h \check{\gamma}_h^{\text{EVB}}. \quad (6.108)$$

Using Eq. (6.108), we obtain the following lemma:

Lemma 6.16 *If and only if*

$$\gamma_h \geq \underline{\gamma}^{\text{local-EVB}} \equiv (\sqrt{L} + \sqrt{M})\sigma, \quad (6.109)$$

the positive EVB local solution given by

$$\widehat{a}_h = \pm \sqrt{\check{\gamma}_h^{\text{EVB}} \widehat{\delta}_h^{\text{EVB}}}, \quad \widehat{b}_h = \pm \sqrt{\frac{\check{\gamma}_h^{\text{EVB}}}{\widehat{\delta}_h^{\text{EVB}}}}, \quad (6.110)$$

$$\widehat{\sigma}_{a_h}^2 = \frac{\sigma^2 \widehat{\delta}_h^{\text{EVB}}}{\gamma_h}, \quad \widehat{\sigma}_{b_h}^2 = \frac{\sigma^2}{\gamma_h \widehat{\delta}_h^{\text{EVB}}}, \quad c_{a_h}c_{b_h} = \sqrt{\frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{LM}}, \quad (6.111)$$

exists with the following free energy:

$$F_h^{\text{EVB-Posi}} = M \log\left(\frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{M\sigma^2} + 1\right) + L \log\left(\frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{L\sigma^2} + 1\right) - \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{\sigma^2}. \quad (6.112)$$

Here,

$$\widehat{\delta}_h^{\text{EVB}} = \sqrt{\frac{M\check{\gamma}_h^{\text{EVB}}}{L\gamma_h}} \left(1 + \frac{L\sigma^2}{\gamma_h \check{\gamma}_h^{\text{EVB}}} \right), \quad (6.113)$$

and $\check{\gamma}_h^{\text{EVB}}$ is given by Eq. (6.103).

Proof Lemma 6.15 immediately leads to the EVB shrinkage estimator (6.103). We can find the value of $c_{a_h} c_{b_h}$ at the positive EVB local solution by combining the condition (6.78) for the VB estimator and the condition (6.108) for the EVB estimator. Specifically, by using the condition (6.108), the condition (6.78) for $\check{\gamma}_h^{\text{VB}}$ replaced with $\check{\gamma}_h^{\text{EVB}}$ can be written as

$$\left(\gamma_h - \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{\check{\gamma}_h^{\text{EVB}} + \frac{M\sigma^2}{\gamma_h}} \right) \left(\gamma_h - \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{\check{\gamma}_h^{\text{EVB}} + \frac{L\sigma^2}{\gamma_h}} \right) = \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2},$$

and therefore

$$\left(\frac{M\sigma^2}{\check{\gamma}_h^{\text{EVB}} + \frac{M\sigma^2}{\gamma_h}} \right) \left(\frac{L\sigma^2}{\check{\gamma}_h^{\text{EVB}} + \frac{L\sigma^2}{\gamma_h}} \right) = \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2}.$$

Applying the condition (6.108) again gives

$$\frac{LM\sigma^4}{\gamma_h \check{\gamma}_h^{\text{EVB}}} = \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2},$$

which leads to the last equation in Eq. (6.111).

Similarly, using the condition (6.108), Eq. (6.52) for $\check{\gamma}_h^{\text{VB}}$ replaced with $\check{\gamma}_h^{\text{EVB}}$ is written as

$$\begin{aligned} \widehat{\delta}_h &= \frac{c_{a_h}^2}{\sigma^2} \left(\gamma_h - \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{\check{\gamma}_h^{\text{EVB}} + \frac{M\sigma^2}{\gamma_h}} \right) \\ &= \frac{c_{a_h}^2}{\sigma^2} \left(\frac{M\sigma^2}{\check{\gamma}_h^{\text{EVB}} + \frac{M\sigma^2}{\gamma_h}} \right) \\ &= \frac{c_{a_h}^2 M}{\gamma_h} \left(\frac{\gamma_h \check{\gamma}_h^{\text{EVB}} + L\sigma^2}{\gamma_h \check{\gamma}_h^{\text{EVB}}} \right) \\ &= \frac{c_{a_h}^2 M}{\gamma_h} \left(1 + \frac{L\sigma^2}{\gamma_h \check{\gamma}_h^{\text{EVB}}} \right). \end{aligned}$$

Using the assumption that $c_{a_h} = c_{b_h}$ and therefore $c_{a_h}^2 = c_{a_h} c_{b_h}$, we obtain Eq. (6.113). Eq. (6.110) and the first two equations in Eq. (6.111) are simply obtained from Lemma 6.10.

Finally, applying Eq. (6.108) to the free energy (6.72), we have

$$F_h^{\text{EVB-Posi}} = -M \log \left(1 - \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{\gamma_h \check{\gamma}_h^{\text{EVB}} + M\sigma^2} \right) - L \log \left(1 - \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{\gamma_h \check{\gamma}_h^{\text{EVB}} + L\sigma^2} \right) - \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{\sigma^2},$$

which leads to Eq. (6.112). This completes the proof of Lemma 6.16. \square

In Figure 6.3, the positive EVB local solution at $c_{a_h} c_{b_h} = \sqrt{\gamma_h \check{\gamma}_h^{\text{EVB}} / (LM)}$ is indicated by a cross if it exists.

6.9.2 EVB Threshold

Lemmas 6.14 and 6.16 state that, if $\gamma_h \leq \underline{\gamma}^{\text{local-EVB}}$, only the null EVB local solution exists, and therefore it is the global EVB solution. In this section, assuming that $\gamma_h \geq \underline{\gamma}^{\text{local-EVB}}$, we compare the free energy (6.105) at the null EVB local solution and the free energy (6.112) at the positive EVB local solution. Since $F_h^{\text{EVB-Null}} \rightarrow +0$, we simply consider the situation where $F_h^{\text{EVB-Posi}} \leq 0$. Eq. (6.108) gives

$$(\gamma_h \check{\gamma}_h^{\text{EVB}} + L\sigma^2) \left(1 + \frac{M\sigma^2}{\gamma_h \check{\gamma}_h^{\text{EVB}}} \right) = \gamma_h^2. \quad (6.114)$$

By using Eqs. (6.103) and (6.109), we have

$$\begin{aligned} \gamma_h \check{\gamma}_h^{\text{EVB}} &= \frac{1}{2} \left(\gamma_h^2 - (\underline{\gamma}^{\text{local-EVB}})^2 + 2\sqrt{LM}\sigma^2 \right. \\ &\quad \left. + \sqrt{(\gamma_h^2 - (\underline{\gamma}^{\text{local-EVB}})^2)(\gamma_h^2 - (\underline{\gamma}^{\text{local-EVB}})^2 + 4\sqrt{LM}\sigma^2)} \right) \\ &\geq \sqrt{LM}\sigma^2. \end{aligned} \quad (6.115)$$

Remember the definition of α (Eq. (6.99))

$$\alpha = \frac{L}{M} \quad (0 < \alpha \leq 1),$$

and let

$$x_h = \frac{\gamma_h^2}{M\sigma^2}, \quad (6.116)$$

$$\tau_h = \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{M\sigma^2}. \quad (6.117)$$

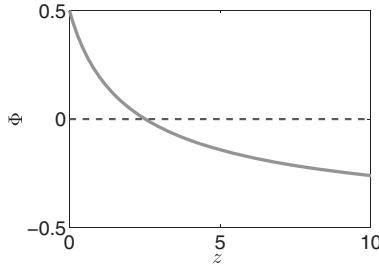


Figure 6.5 $\Phi(z) = \frac{\log(z+1)}{z} - \frac{1}{2}$.

Eqs. (6.114) and (6.103) imply the following mutual relations between x_h and τ_h :

$$x_h \equiv x(\tau_h; \alpha) = (1 + \tau_h) \left(1 + \frac{\alpha}{\tau_h} \right), \quad (6.118)$$

$$\tau_h \equiv \tau(x_h; \alpha) = \frac{1}{2} \left(x_h - (1 + \alpha) + \sqrt{(x_h - (1 + \alpha))^2 - 4\alpha} \right). \quad (6.119)$$

Eqs. (6.109) and (6.115) lead to

$$x_h \geq \underline{x}^{\text{local}} = \frac{(\gamma^{\text{local-EVB}})^2}{M\sigma^2} = x(\sqrt{\alpha}; \alpha) = (1 + \sqrt{\alpha})^2, \quad (6.120)$$

$$\tau_h \geq \underline{\tau}^{\text{local}} = \sqrt{\alpha}. \quad (6.121)$$

Then, using $\Xi(\tau; \alpha)$ defined by Eq. (6.100), we can rewrite Eq. (6.112) as

$$\begin{aligned} F_h^{\text{EVB-Posi}} &= M \log(\tau_h + 1) + L \log\left(\frac{\tau_h}{\alpha} + 1\right) - M\tau_h \\ &= M\tau_h \Xi(\tau; \alpha). \end{aligned} \quad (6.122)$$

The following holds for $\Phi(z)$ (which is also defined in Eq. (6.100)):

Lemma 6.17 $\Phi(z)$ is decreasing for $z > 0$.

Proof The derivative is

$$\frac{\partial \Phi}{\partial z} = \frac{1 - \frac{1}{z+1} - \log(z+1)}{z^2},$$

which is negative for $z > 0$ because

$$\frac{1}{z+1} + \log(z+1) > 1.$$

This completes the proof of Lemma 6.17. \square

Figure 6.5 shows the profile of $\Phi(z)$. Since $\Phi(z)$ is decreasing, $\Xi(\tau; \alpha)$ is also decreasing with respect to τ . It holds that, for any $0 < \alpha \leq 1$,

$$\lim_{\tau \rightarrow 0} \Xi(\tau; \alpha) = 1,$$

$$\lim_{\tau \rightarrow \infty} \Xi(\tau; \alpha) = -1.$$

Therefore, $\Xi(\tau; \alpha)$ has a unique zero-cross point $\underline{\tau}$, such that

$$\Xi(\tau; \alpha) \leq 0 \quad \text{if and only if} \quad \tau \geq \underline{\tau}. \quad (6.123)$$

Then, we can prove the following lemma:

Lemma 6.18 *The unique zero-cross point $\underline{\tau}$ of $\Xi(\tau; \alpha)$ lies in the following range:*

$$\sqrt{\alpha} < \underline{\tau} < \underline{z},$$

where $\underline{z} \approx 2.5129$ is the unique zero-cross point of $\Phi(z)$.

Proof Since $\Phi(z)$ is decreasing, $\Xi(\tau; \alpha)$ is upper-bounded by

$$\Xi(\tau; \alpha) = \Phi(\tau) + \Phi\left(\frac{\tau}{\alpha}\right) \leq 2\Phi(\tau) = \Xi(\tau; 1).$$

Therefore, the unique zero-cross point $\underline{\tau}$ of $\Xi(\tau; \alpha)$ is no greater than the unique zero-cross point \underline{z} of $\Phi(z)$, i.e.,

$$\underline{\tau} \leq \underline{z}.$$

For obtaining the lower-bound $\underline{\tau} > \sqrt{\alpha}$, it suffices to show that $\Xi(\sqrt{\alpha}; \alpha) > 0$. Let us prove that the following function is decreasing and positive for $0 < \alpha \leq 1$:

$$g(\alpha) \equiv \frac{\Xi(\sqrt{\alpha}; \alpha)}{\sqrt{\alpha}}.$$

From the definition (6.100) of $\Xi(\tau; \alpha)$, we have

$$g(\alpha) = \left(1 + \frac{1}{\alpha}\right) \log(\sqrt{\alpha} + 1) - \log \sqrt{\alpha} - \frac{1}{\sqrt{\alpha}}.$$

The derivative is given by

$$\begin{aligned} \frac{\partial g}{\partial \sqrt{\alpha}} &= \frac{\left(1 + \frac{1}{\alpha}\right)}{\sqrt{\alpha} + 1} - \frac{2}{\alpha^{3/2}} \log(\sqrt{\alpha} + 1) - \frac{1}{\sqrt{\alpha}} + \frac{1}{\alpha} \\ &= -\frac{2}{\alpha^{3/2}} \left(\log(\sqrt{\alpha} + 1) + \frac{1}{\sqrt{\alpha} + 1} - 1 \right) \\ &< 0, \end{aligned}$$

which implies that $g(\alpha)$ is decreasing. Since

$$g(1) = 2 \log 2 - 1 \approx 0.3863 > 0,$$

$g(\alpha)$ is positive for $0 < \alpha \leq 1$, which completes the proof of Lemma 6.18. \square

Since Eq. (6.118) is increasing with respect to τ_h ($> \sqrt{\alpha}$), the thresholding condition $\tau \geq \underline{\tau}$ in Eq. (6.123) can be expressed in terms of x :

$$\Xi(\tau(x); \alpha) \leq 0 \quad \text{if and only if} \quad x \geq \underline{x}, \quad (6.124)$$

$$\text{where} \quad \underline{x} \equiv x(\underline{\tau}; \alpha) = \left(1 + \frac{\underline{\tau}}{\underline{\tau}}\right) \left(1 + \frac{\alpha}{\underline{\tau}}\right). \quad (6.125)$$

Using Eqs. (6.116) and (6.122), we have

$$F_h^{\text{EVB-Posi}} \leq 0 \quad \text{if and only if} \quad \gamma_h \geq \underline{\gamma}^{\text{EVB}}, \quad (6.126)$$

where $\underline{\gamma}^{\text{EVB}}$ is defined by Eq. (6.102). Thus, we have the following lemma:

Lemma 6.19 *The positive EVB local solution is the global EVB solution if and only if $\gamma_h \geq \underline{\gamma}^{\text{EVB}}$.*

Combining Lemmas 6.14, 6.16, and 6.19 completes the proof of Theorem 6.13. \square

Figure 6.6 shows estimators and thresholds for $L = M = H = 1$ and $\sigma^2 = 1$. The curves indicate the VB solution $\widehat{\gamma}_h^{\text{VB}}$ given by Eq. (6.48), the EVB solution $\widehat{\gamma}_h^{\text{EVB}}$ given by Eq. (6.101), the EVB positive local minimizer $\check{\gamma}_h^{\text{EVB}}$ given by Eq. (6.103), and the EVB positive local maximizer $\dot{\gamma}_h$ given by Eq. (6.107), respectively. The arrows indicate the VB threshold $\underline{\gamma}_h^{\text{VB}}$ given by Eq. (6.49), the local EVB threshold $\underline{\gamma}^{\text{local-EVB}}$ given by Eq. (6.109), and the EVB threshold $\underline{\gamma}^{\text{EVB}}$ given by Eq. (6.102), respectively.

6.10 Summary of Intermediate Results

In the rest of this section, we summarize some intermediate results obtained in Section 6.9, which are useful for further analysis (mainly in Chapter 8).

Summarizing Eqs. (6.109), (6.114), and (6.115) leads to the following corollary:

Corollary 6.20 *The EVB shrinkage estimator (6.103) is a stationary point of the free energy (6.43), which exists if and only if*

$$\gamma_h \geq \underline{\gamma}^{\text{local-EVB}} \equiv (\sqrt{L} + \sqrt{M})\sigma, \quad (6.127)$$

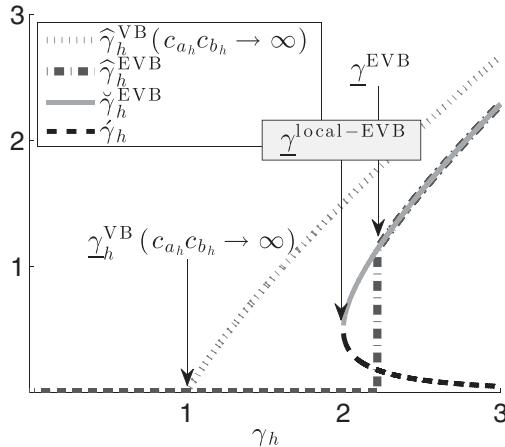


Figure 6.6 Estimators and thresholds for $L = M = H = 1$ and $\sigma^2 = 1$.

and satisfies the following equation:

$$\left(\gamma_h \check{\gamma}_h^{\text{EVB}} + L\sigma^2 \right) \left(1 + \frac{M\sigma^2}{\gamma_h \check{\gamma}_h^{\text{EVB}}} \right) = \gamma_h^2. \quad (6.128)$$

It holds that

$$\gamma_h \check{\gamma}_h^{\text{EVB}} \geq \sqrt{LM}\sigma^2. \quad (6.129)$$

Combining Lemmas 6.14, 6.16, and 6.19 leads to the following corollary:

Corollary 6.21 *The minimum free energy achieved under EVB learning is given by Eq. (6.42) with*

$$2F_h = \begin{cases} M \log \left(\frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{M\sigma^2} + 1 \right) + L \log \left(\frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{L\sigma^2} + 1 \right) - \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{\sigma^2} & \text{if } \gamma_h \geq \underline{\gamma}^{\text{EVB}}, \\ +0 & \text{otherwise.} \end{cases} \quad (6.130)$$

Corollary 6.20 together with Theorem 6.13 implies that when

$$\underline{\gamma}^{\text{local-EVB}} \leq \gamma_h < \underline{\gamma}^{\text{EVB}},$$

a stationary point (called the *positive* EVB local solution and specified by Lemma 6.16) exists at Eq. (6.103), but it is not the global minimum. Actually, a local minimum (called the *null* EVB local solution and specified by Lemma 6.14) with $F_h = +0$ always exists. The stationary point at Eq. (6.103) is a *nonglobal* local minimum when $\underline{\gamma}^{\text{local-EVB}} \leq \gamma_h < \underline{\gamma}^{\text{EVB}}$ and the global

minimum when $\gamma_h \geq \underline{\gamma}^{\text{EVB}}$ (see Figure 6.3 with its caption). This phase transition induces the free energy thresholding observed in Corollary 6.21.

We define a *local-EVB estimator* by

$$\widehat{U}^{\text{local-EVB}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{local-EVB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top,$$

where $\widehat{\gamma}_h^{\text{local-EVB}} = \begin{cases} \check{\gamma}_h^{\text{EVB}} & \text{if } \gamma_h \geq \underline{\gamma}^{\text{local-EVB}}, \\ 0 & \text{otherwise,} \end{cases}$

(6.131)

and call $\underline{\gamma}^{\text{local-EVB}}$ a local-EVB threshold. This estimator gives the positive EVB local solution, whenever it exists, for each singular component. There is an interesting relation between the *local-EVB* solution and an alternative dimensionality selection method (Hoyle, 2008), which will be discussed in Section 8.6.

Rescaling the quantities related to the squared singular value by $M\sigma^2$ —to which the contribution from noise (each eigenvalue of $\mathcal{E}^\top \mathcal{E}$) scales linearly—simplifies expressions. Assume that the condition (6.127) holds, and define

$$x_h = \frac{\gamma_h^2}{M\sigma^2},$$
(6.132)

$$\tau_h = \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{M\sigma^2},$$
(6.133)

which are used as a rescaled observation and a rescaled EVB estimator, respectively. Eqs. (6.128) and (6.103) specify the mutual relations between them:

$$x_h \equiv x(\tau_h; \alpha) = (1 + \tau_h) \left(1 + \frac{\alpha}{\tau_h} \right),$$
(6.134)

$$\tau_h \equiv \tau(x_h; \alpha) = \frac{1}{2} \left(x_h - (1 + \alpha) + \sqrt{(x_h - (1 + \alpha))^2 - 4\alpha} \right).$$
(6.135)

With these rescaled variables, the condition (6.127), as well as (6.129), for the existence of the positive local-EVB solution $\check{\gamma}_h^{\text{EVB}}$ is expressed as

$$x_h \geq \underline{x}^{\text{local}} = \frac{(\underline{\gamma}^{\text{local-EVB}})^2}{M\sigma^2} = x(\sqrt{\alpha}; \alpha) = (1 + \sqrt{\alpha})^2,$$
(6.136)

$$\tau_h \geq \underline{\tau}^{\text{local}} = \sqrt{\alpha}.$$
(6.137)

The EVB threshold (6.102) is expressed as

$$\underline{x} = \frac{(\underline{\gamma}^{\text{EVB}})^2}{M\sigma^2} = x(\underline{\tau}; \alpha) = \left(1 + \underline{\tau} \right) \left(1 + \frac{\alpha}{\underline{\tau}} \right),$$
(6.138)

and the free energy (6.130) is expressed as

$$F_h = M\tau_h \cdot \min(0, \Xi(\tau_h; \alpha)), \quad (6.139)$$

where $\Xi(\tau; \alpha)$ is defined by Eq. (6.100).

The preceding rescaled expressions give an intuition of Theorem 6.13: the EVB solution $\widehat{\gamma}_h^{\text{EVB}}$ is positive if and only if the positive local-EVB solution $\check{\gamma}_h^{\text{EVB}}$ exists (i.e., $x_h \geq \underline{x}^{\text{local}}$), and the free energy $\Xi(\tau(x_h; \alpha); \alpha)$ at the local-EVB solution is nonpositive (i.e., $\tau(x_h; \alpha) \geq \underline{\tau}$ or equivalently $x_h \geq \underline{x}$).

7

Model-Induced Regularization and Sparsity Inducing Mechanism

Variational Bayesian (VB) learning often shows the automatic relevance determination (ARD) property—the solution is sparse with unnecessary components eliminated automatically. In this chapter, we try to elucidate the sparsity inducing mechanism of VB learning, based on the global analytic solutions derived in Chapter 6. We argue that the ARD property is induced by the model-induced regularization (MIR), which all Bayesian learning methods possess when *unidentifiable* models are involved, and that MIR is enhanced by the independence constraint (imposed for computational tractability), which induces phase transitions making the solution (exactly) sparse (Nakajima and Sugiyama, 2011).

We first show the VB solution for special cases where the MIR effect is visible in the solution form. Then we illustrate the behavior of the posteriors and estimators in the one-dimensional case, comparing VB learning with maximum a posteriori (MAP) learning and Bayesian learning. After that, we explain MIR, and how it is enhanced in VB learning through phase transitions.

7.1 VB Solutions for Special Cases

Here we discuss two special cases of fully observed matrix factorization (MF), in which the VB solution is simple and intuitive.

Almost Flat Prior

When $c_{a_h} c_{b_h} \rightarrow \infty$ (i.e., the prior is *almost flat*), the VB solution given by Theorem 6.7 in Chapter 6 has a simple form.

Corollary 7.1 *The VB solution of the fully observed matrix factorization model (6.1) through (6.3) is given by*

$$\widehat{\mathbf{U}}^{\text{VB}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \quad (7.1)$$

where the estimator $\widehat{\gamma}_h^{\text{VB}}$ corresponding to the h th largest singular value is upper-bounded as

$$\widehat{\gamma}_h^{\text{VB}} < \max \left\{ 0, \left(1 - \frac{\max(L, M)\sigma^2}{\gamma_h^2} \right) \gamma_h \right\}. \quad (7.2)$$

For the almost flat prior (i.e., $c_{a_h} c_{b_h} \rightarrow \infty$), the equality holds, i.e.,

$$\lim_{c_{a_h} c_{b_h} \rightarrow \infty} \widehat{\gamma}_h^{\text{VB}} = \max \left\{ 0, \left(1 - \frac{\max(L, M)\sigma^2}{\gamma_h^2} \right) \gamma_h \right\}. \quad (7.3)$$

Proof It is clear that the threshold (6.49) is decreasing and the shrinkage factor (6.50) is increasing with respect to $c_{a_h} c_{b_h}$. Therefore, $\widehat{\gamma}_h^{\text{VB}}$ is largest for $c_{a_h} c_{b_h} \rightarrow \infty$. In this limit, Eqs. (6.49) and (6.50) are reduced to

$$\begin{aligned} \lim_{c_{a_h} c_{b_h} \rightarrow \infty} \widehat{\gamma}_h^{\text{VB}} &= \sigma \sqrt{\frac{(L+M)}{2} + \sqrt{\left(\frac{(L+M)}{2}\right)^2 - LM}} \\ &= \sigma \sqrt{\max(L, M)}, \\ \lim_{c_{a_h} c_{b_h} \rightarrow \infty} \check{\gamma}_h^{\text{VB}} &= \gamma_h \left(1 - \frac{\sigma^2}{2\gamma_h^2} \left(M + L + \sqrt{(M-L)^2} \right) \right) \\ &= \left(1 - \frac{\max(L, M)\sigma^2}{\gamma_h^2} \right) \gamma_h, \end{aligned}$$

which prove the corollary. \square

The form of the VB solution (7.3) in the limit is known as the *positive-part James–Stein (PJS) estimator* (James and Stein, 1961), operated on each singular component separately (see Appendix A for its interesting property and the relation to Bayesian learning). A counterintuitive fact—a shrinkage is observed even in the limit of the flat prior—will be explained in terms of MIR in Section 7.3.

Square Matrix

When $L = M$ (i.e., the observed matrix V is square), the VB solution is intuitive, so that the shrinkage caused by MIR and the shrinkage caused by the prior are separately visible in its formula.

Corollary 7.2 When $L = M$, the VB solution is given by Eq. (7.1) with

$$\hat{\gamma}_h^{\text{VB}} = \max \left\{ 0, \left(1 - \frac{M\sigma^2}{\gamma_h^2} \right) \gamma_h - \frac{\sigma^2}{c_{a_h} c_{b_h}} \right\}. \quad (7.4)$$

Proof When $L = M$, Eqs. (6.49) and (6.50) can be written as

$$\begin{aligned} \underline{\gamma}_h^{\text{VB}} &= \sigma \sqrt{M + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2} + \sqrt{\left(M + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2} \right)^2 - M^2}}, \\ \check{\gamma}_h^{\text{VB}} &= \gamma_h \left(1 - \frac{\sigma^2}{2\gamma_h^2} \left(2M + \sqrt{\frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) \right) \\ &= \gamma_h \left(1 - \frac{\sigma^2}{\gamma_h^2} \left(M + \frac{\gamma_h}{c_{a_h} c_{b_h}} \right) \right). \end{aligned}$$

We can confirm that $\check{\gamma}_h^{\text{VB}} \leq 0$ when $\gamma_h \leq \underline{\gamma}_h^{\text{VB}}$, which proves the corollary. Actually, we can confirm that $\check{\gamma}_h^{\text{VB}} = 0$ when $\gamma_h = \underline{\gamma}_h^{\text{VB}}$, and $\check{\gamma}_h^{\text{VB}} < 0$ when $\gamma_h < \underline{\gamma}_h^{\text{VB}}$ for any $L, M, c_{a_h}^2$, and $c_{b_h}^2$. \square

In the VB solution (7.4), we can identify the PJS shrinkage and a constant shrinkage. The PJS shrinkage can be considered to be caused by MIR since it appears even with the flat prior, while the constant shrinkage $-\sigma^2/(c_{a_h} c_{b_h})$ is considered to be caused by the prior since it appears in MAP learning (see Theorem 12.1 in Chapter 12).

The empirical VB (EVB) solution is also simple for square matrices. The following corollary is obtained from Theorem 6.13 in Chapter 6:

Corollary 7.3 When $L = M$, the global EVB solution is given by

$$\hat{\gamma}_h^{\text{EVB}} = \begin{cases} \check{\gamma}_h^{\text{EVB}} & \text{if } \gamma_h > \underline{\gamma}_h^{\text{EVB}}, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\begin{aligned} \underline{\gamma}_h^{\text{EVB}} &= \sigma \sqrt{M \left(2 + \underline{\tau}(1) + \frac{1}{\underline{\tau}(1)} \right)}, \\ \check{\gamma}_h^{\text{EVB}} &= \frac{\gamma_h}{2} \left(1 - \frac{2M\sigma^2}{\gamma_h^2} + \sqrt{1 - \frac{4M\sigma^2}{\gamma_h^2}} \right). \end{aligned}$$

Proof When $L = M$, Eqs. (6.102) and (6.103) can be written as

$$\begin{aligned}\underline{\gamma}^{\text{EVB}} &= \sigma \sqrt{M \left(2 + \underline{\tau}(1) + \frac{1}{\underline{\tau}(1)} \right)}, \\ \check{\gamma}_h^{\text{EVB}} &= \frac{\gamma_h}{2} \left(1 - \frac{2M\sigma^2}{\gamma_h^2} + \sqrt{\left(1 - \frac{2M\sigma^2}{\gamma_h^2} \right)^2 - \frac{4M^2\sigma^4}{\gamma_h^4}} \right) \\ &= \frac{\gamma_h}{2} \left(1 - \frac{2M\sigma^2}{\gamma_h^2} + \sqrt{1 - \frac{4M\sigma^2}{\gamma_h^2}} \right),\end{aligned}$$

which completes the proof. \square

7.2 Posteriors and Estimators in a One-Dimensional Case

In order to illustrate how strongly Bayesian learning and its approximation methods are regularized, we depict posteriors and estimators in the MF model for $L = M = H = 1$ (i.e., U , V , A , and B are merely scalars):

$$p(V|A, B) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(V - BA)^2}{2\sigma^2}\right). \quad (7.5)$$

In this model, we can visualize the *unidentifiability* of the MF model as *equivalence classes*—a set of points (A, B) on which the product is unchanged, i.e., $U = BA$, represents the same distribution (see Figure 7.1). When $U = 0$, the equivalence class has a “cross-shape” profile on the A - and B -axes; otherwise, it forms a pair of hyperbolic curves. This redundant structure in the

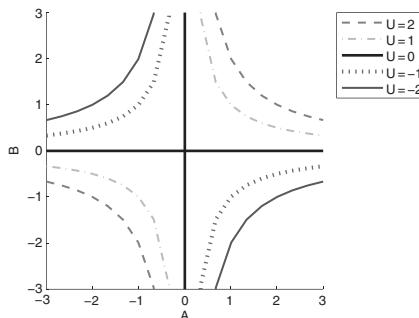


Figure 7.1 Equivalence class structure of the one-dimensional MF model. Any A and B such that their product is unchanged give the same U .

parameter space is the origin of MIR, and highly influences the phase transition phenomenon in VB learning, as we will see shortly.

With Gaussian priors,

$$p(A) = \frac{1}{\sqrt{2\pi c_a^2}} \exp\left(-\frac{A^2}{2c_a^2}\right), \quad (7.6)$$

$$p(B) = \frac{1}{\sqrt{2\pi c_b^2}} \exp\left(-\frac{B^2}{2c_b^2}\right), \quad (7.7)$$

the Bayes posterior is proportional to

$$\begin{aligned} p(A, B|V) &\propto p(V|A, B)p(A)p(B) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(V - BA)^2 - \frac{A^2}{2c_a^2} - \frac{B^2}{2c_b^2}\right). \end{aligned} \quad (7.8)$$

Figure 7.2 shows the contour of the *unnormalized* Bayes posterior (7.8) when $V = 0, 1, 2$ are observed, the noise variance is $\sigma^2 = 1$, and the prior covariances are set to $c_a = c_b = 100$ (i.e., almost flat priors). We can see that the equivalence class structure is reflected in the Bayes posterior: when $V = 0$, the surface of the Bayes posterior has a cross-shaped profile and its maximum is at the origin; when $V > 0$, the surface is divided into the positive orthant (i.e., $A, B > 0$) and the negative orthant (i.e., $A, B < 0$), and the two “modes” get farther as V increases.

MAP Solution

Let us first investigate the behavior of the *MAP estimator*, which coincides with the maximum likelihood (ML) estimator when the priors are flat. For finite c_a and c_b , the MAP solution can be expressed as

$$\begin{aligned} \widehat{A}^{\text{MAP}} &= \pm \sqrt{\frac{c_a}{c_b} \max\left\{0, |V| - \frac{\sigma^2}{c_a c_b}\right\}}, \\ \widehat{B}^{\text{MAP}} &= \pm \text{sign}(V) \sqrt{\frac{c_b}{c_a} \max\left\{0, |V| - \frac{\sigma^2}{c_a c_b}\right\}}, \end{aligned}$$

where $\text{sign}(\cdot)$ denotes the sign of a scalar (see Corollary 12.2 in Chapter 12 for derivation). In Figure 7.2, the asterisks indicate the MAP estimators, and the dashed curves indicate the ML estimators (the modes of the contour of Eq. (7.8) when $c_a = c_b \rightarrow \infty$). When $V = 0$, the Bayes posterior takes the maximum value on the A - and B -axes, which results in the MAP estimator equal to $\widehat{U}^{\text{MAP}} (= \widehat{B}^{\text{MAP}} \widehat{A}^{\text{MAP}}) = 0$. When $V = 1$, the profile of the Bayes posterior is hyperbolic and the maximum value is achieved on the hyperbolic curves in the positive orthant (i.e., $A, B > 0$) and the negative orthant (i.e., $A, B < 0$);

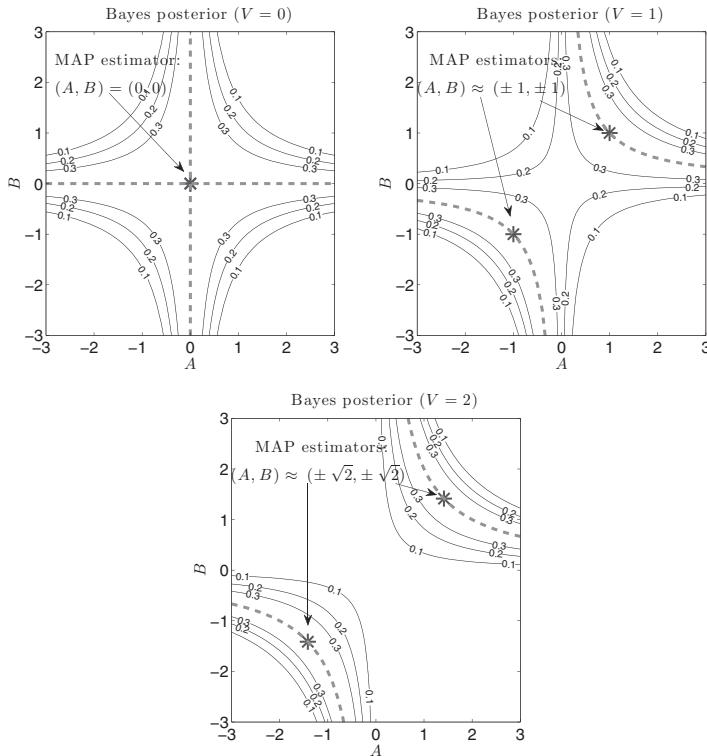


Figure 7.2 (Unnormalized) Bayes posteriors for $c_a = c_b = 100$ (i.e., almost flat priors). The asterisks are the MAP estimators, and the dashed curves indicate the ML estimators (the modes of the contour when $c_a = c_b = c \rightarrow \infty$).

in either case, $\widehat{U}^{\text{MAP}} \approx 1$ ($\lim_{c_a, c_b \rightarrow \infty} \widehat{U}^{\text{MAP}} = 1$). When $V = 2$, a similar multimodal structure is observed and the MAP estimator is $\widehat{U}^{\text{MAP}} \approx 2$ ($\lim_{c_a, c_b \rightarrow \infty} \widehat{U}^{\text{MAP}} = 2$). From these plots, we can visually confirm that the MAP estimator with almost flat priors ($c_a = c_b = 100$) approximately agrees with the ML estimator: $\widehat{U}^{\text{MAP}} \approx \widehat{U}^{\text{ML}} = V$ ($\lim_{c_a, c_b \rightarrow \infty} \widehat{U}^{\text{MAP}} = \widehat{U}^{\text{ML}}$). We will use the ML estimator as an *unregularized* reference in the following discussion.

Figure 7.3 shows the contour of the Bayes posterior when $c_a = c_b = 2$. The MAP estimators shift from the ML solutions (dashed curves) toward the origin, and they are more clearly contoured as peaks.

VB Solution

Next we depict the VB posterior, given by Corollary 6.8 in Chapter 6. When $L = M = H = 1$, the VB solution is given by

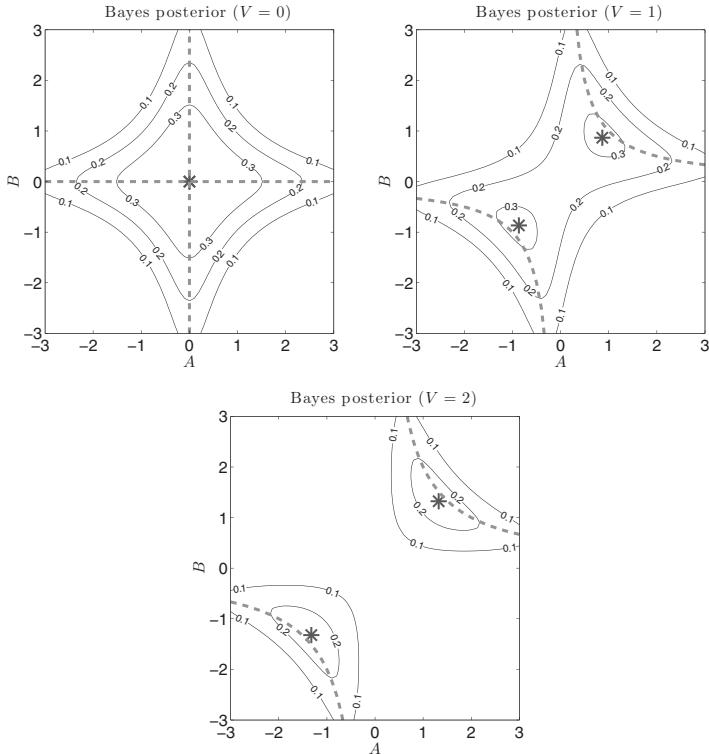


Figure 7.3 (Unnormalized) Bayes posteriors for $c_a = c_b = 2$. The dashed curves indicating the ML estimators are identical to those in Figure 7.2.

$$r(A, B) = \begin{cases} \text{Gauss}_1\left(A; \pm \sqrt{\check{\gamma}^{\text{VB}} \frac{c_a}{c_b}}, \frac{\sigma^2 c_a}{|V| c_b}\right) \text{Gauss}_1\left(B; \pm \text{sign}(V) \sqrt{\check{\gamma}^{\text{VB}} \frac{c_b}{c_a}}, \frac{\sigma^2 c_b}{|V| c_a}\right) & \text{if } |V| \geq \underline{\gamma}^{\text{VB}}, \\ \text{Gauss}_1\left(A; 0, c_a^2 \check{\kappa}^{\text{VB}}\right) \text{Gauss}_1\left(B; 0, c_b^2 \check{\kappa}^{\text{VB}}\right) & \text{otherwise,} \end{cases} \quad (7.9)$$

where

$$\begin{aligned} \underline{\gamma}^{\text{VB}} &= \sigma \sqrt{1 + \frac{\sigma^2}{2c_a^2 c_b^2} + \sqrt{\left(1 + \frac{\sigma^2}{2c_a^2 c_b^2}\right)^2 - 1}}, \\ \check{\gamma}^{\text{VB}} &= \left(1 - \frac{\sigma^2}{V^2}\right) |V| - \frac{\sigma^2}{c_a c_b}, \\ \check{\kappa}^{\text{VB}} &= -\frac{\sigma^2}{2c_a^2 c_b^2} + \sqrt{\left(1 + \frac{\sigma^2}{2c_a^2 c_b^2}\right)^2 - 1}. \end{aligned}$$

Figure 7.4 shows the contour of the VB posterior (7.9) when $V = 0, 1, 2$ are observed, the noise variance is $\sigma^2 = 1$, and the prior covariances are

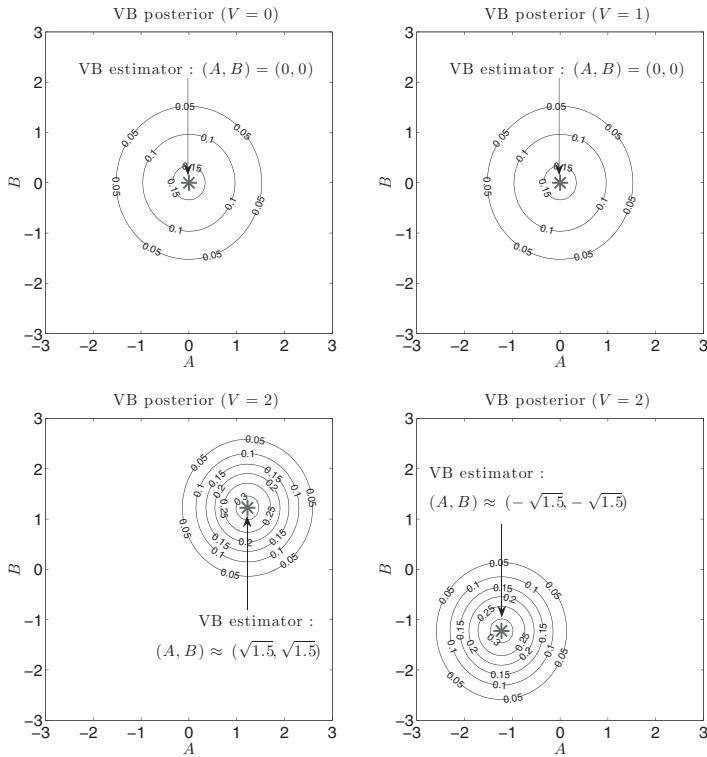


Figure 7.4 VB solutions for $c_a = c_b = 100$ (i.e., almost flat priors). When $V = 2$, VB learning gives either one of the two solutions shown in the bottom row.

set to $c_a = c_b = 100$ (i.e., almost flat priors). When $V = 0$, the cross-shaped contour of the Bayes posterior (see Figure 7.2) is approximated by a spherical Gaussian distribution located at the origin. Thus, the VB estimator is $\widehat{U}^{\text{VB}} = 0$, which coincides with the MAP estimator. When $V = 1$, two hyperbolic “modes” of the Bayes posterior are approximated again by a spherical Gaussian distribution located at the origin. Thus, the VB estimator is still $\widehat{U}^{\text{VB}} = 0$, which differs from the MAP estimator $\widehat{U}^{\text{MAP}} \approx 1$.

$V = \underline{\gamma}^{\text{VB}} \approx \sqrt{M\sigma^2} = 1$ ($\lim_{c_a, c_b \rightarrow \infty} \underline{\gamma}^{\text{VB}} = \sqrt{M\sigma^2}$) is actually a transition point of the VB solution. When V is not larger than the threshold $\underline{\gamma}^{\text{VB}} \approx 1$, VB learning tries to approximate the two “modes” of the Bayes posterior by the origin-centered Gaussian distribution. When V goes beyond the threshold $\underline{\gamma}^{\text{VB}} \approx 1$, the “distance” between two hyperbolic modes of the Bayes posterior becomes so large that VB learning chooses to approximate one of those two modes in the positive and the negative orthants. As such, the symmetry is

broken spontaneously and the VB estimator is detached from the origin. The bottom row of Figure 7.4 shows the contour of the two possible VB posteriors when $V = 2$. Note that the VB estimator, $\widehat{U}^{\text{VB}} \approx 3/2$, is the same for both cases, and differs from the MAP estimator $\widehat{U}^{\text{MAP}} \approx 2$.

In general, the VB estimator is closer to the origin than the MAP estimator, and the relative difference between them tends to shrink as V increases.

Bayesian Estimator

The full Bayesian estimator is defined as the mean of the Bayes posterior (see Eq. (1.7)). In the MF model with $L = M = H = 1$, the Bayesian estimator is expressed as

$$\widehat{U}^{\text{Bayes}} = \langle BA \rangle_{p(V|A,B)p(A)p(B)/p(V)}. \quad (7.10)$$

If $V = 0, 1, 2, 3$ are observed, the Bayesian estimator with almost flat priors are $\widehat{U}^{\text{Bayes}} = 0, 0.92, 1.93, 2.95$, respectively, which were numerically computed.¹ Compared with the MAP estimator (with almost flat priors), which gives $\widehat{U}^{\text{MAP}} = 0, 1, 2, 3$, respectively, the Bayesian estimator is slightly shrunken.

EVB Solution

Next we consider the empirical Bayesian solutions, where the hyperparameters c_a, c_b are also estimated from observation (the noise variance σ^2 is still treated as a given constant). We fix the ratio between the prior variances to $c_a/c_b = 1$.

From Corollary 7.3 and Eq. (7.9), we obtain the EVB posterior for $L = M = H = 1$ as follows:

$$r(A, B) = \begin{cases} \text{Gauss}_1 \left(A; \pm \sqrt{\check{\gamma}^{\text{EVB}}}, \frac{\sigma^2}{|V|} \right) \text{Gauss}_1 \left(B; \pm \text{sign}(V) \sqrt{\check{\gamma}^{\text{EVB}}}, \frac{\sigma^2}{|V|} \right) & \text{if } |V| \geq \underline{\gamma}^{\text{EVB}}, \\ \text{Gauss}_1(A; 0, +0) \text{Gauss}_1(B; 0, +0) & \text{otherwise,} \end{cases} \quad (7.11)$$

where

$$\begin{aligned} \underline{\gamma}^{\text{EVB}} &= \sigma \sqrt{2 + \underline{\tau}(1) + \frac{1}{\underline{\tau}(1)}} \approx \sigma \sqrt{2 + 2.5129 + \frac{1}{2.5129}} \approx 2.216\sigma, \\ \check{\gamma}^{\text{EVB}} &= \frac{|V|}{2} \left(1 - \frac{2\sigma^2}{V^2} + \sqrt{1 - \frac{4\sigma^2}{V^2}} \right). \end{aligned}$$

¹ More precisely, we numerically calculated the Bayesian estimator (7.10) by sampling A and B from the almost flat priors $p(A)p(B)$ for $c_a = c_b = 100$ and computing the ratio between the sample averages of $BA \cdot p(V|A, B)$ and $p(V|A, B)$.

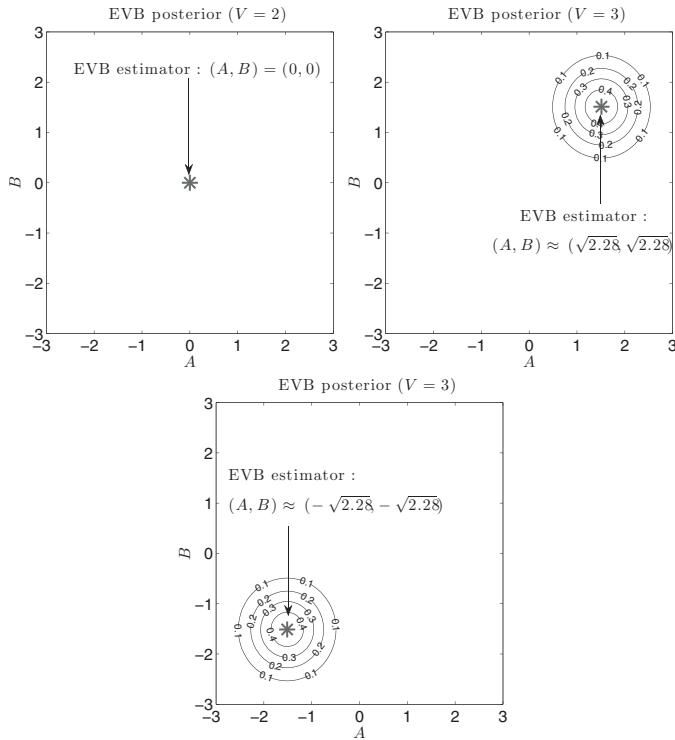


Figure 7.5 EVB solutions. Top-left: When $V = 2$, the EVB posterior is reduced to the Dirac delta function located at the origin. Top-right and bottom: When $V = 3$, the EVB posterior is detached from the origin, and located at $(A, B) \approx (\sqrt{2.28}, \sqrt{2.28})$ or $(A, B) \approx (-\sqrt{2.28}, -\sqrt{2.28})$, both of which yield the same EVB estimator $\widehat{U}^{\text{EVB}} \approx 2.28$.

Figure 7.5 shows the EVB posterior when $V = 2, 3$ are observed, and the noise variance is $\sigma^2 = 1$. When $V = 2 < \gamma^{\text{EVB}}$, the EVB posterior is given by the Dirac delta function located at the origin, resulting in the EVB estimator equal to $\widehat{U}^{\text{EVB}} = 0$ (top-left graph). On the other hand, when $V = 3 > \gamma^{\text{EVB}}$, the EVB posterior is a Gaussian located in the top-right region or bottom-left region, and the EVB estimator is $\widehat{U}^{\text{EVB}} \approx 2.28$ for both solutions (top-right and bottom graphs).

Empirical Bayesian Estimator

The *empirical Bayesian (EBayes) estimator* (introduced in Section 1.2.7) is the Bayesian estimator,

$$\widehat{U}^{\text{EBayes}} = \langle BA \rangle_{p(V|A,B)p(A;\widehat{c}_a)p(B;\widehat{c}_b)/p(V;\widehat{c}_a,\widehat{c}_b)},$$

with the hyperparameters estimated by minimizing the *Bayes free energy* $F^{\text{Bayes}}(V; c_a, c_b) \equiv -\log p(V; c_a, c_b)$, i.e.,

$$(\widehat{c}_a, \widehat{c}_b) = \underset{(c_a, c_b)}{\operatorname{argmin}} F^{\text{Bayes}}(V; c_a, c_b).$$

When $V = 0, 1, 2, 3$ are observed, the EBayes estimators are 0.00, 0.00, 1.25, 2.58 (with the prior variance estimators given by $\widehat{c}_a = \widehat{c}_b \approx 0.0, 0.0, 1.4, 2.1$), respectively, which were numerically computed.²

Behavior of Estimators

Figure 7.6 shows the behavior of estimators, including the MAP estimator \widehat{U}^{MAP} , the VB estimator \widehat{U}^{VB} , the Bayesian estimator $\widehat{U}^{\text{Bayes}}$, the EVB estimator \widehat{U}^{EVB} , and the EBayes estimator $\widehat{U}^{\text{EBayes}}$, when the noise variance is $\sigma^2 = 1$. For nonempirical Bayesian estimators, i.e., the MAP, the VB, and the Bayesian estimators, the hyperparameters are set to $c_a = c_b = 100$ (i.e., almost flat priors). Overall, the solutions satisfy

$$\widehat{U}^{\text{EVB}} < \widehat{U}^{\text{EBayes}} < \widehat{U}^{\text{VB}} < \widehat{U}^{\text{Bayes}} < \widehat{U}^{\text{MAP}} (\approx \widehat{U}^{\text{ML}}),$$

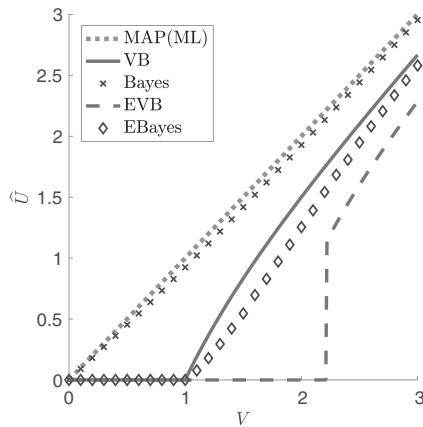


Figure 7.6 Behavior of the MAP estimator \widehat{U}^{MAP} , the VB estimator \widehat{U}^{VB} , the Bayesian estimator $\widehat{U}^{\text{Bayes}}$, the EVB estimator \widehat{U}^{EVB} , and the EBayes estimator $\widehat{U}^{\text{EBayes}}$, when the noise variance is $\sigma^2 = 1$. For the MAP, the VB, and the Bayesian estimators, the hyperparameters are set to $c_a = c_b = 100$ (i.e., almost flat priors).

² For $c_a c_b = 10^{-2.00}, 10^{-1.99}, \dots, 10^{1.00}$, we numerically computed the Bayes free energy, and chose its minimizer $\widehat{c}_a \widehat{c}_b$, with which the Bayesian estimator was computed.

which shows the strength of the regularization effect of each method. Naturally, the empirical Bayesian variants are more regularized than their nonempirical Bayesian counterparts with almost flat priors.

With almost flat priors, the MAP estimator is almost identical to the ML estimator, $\widehat{U}^{\text{MAP}} \approx \widehat{U}^{\text{ML}} = V$, meaning that it is unregularized. We see in Figure 7.6 that the Bayesian estimator $\widehat{U}^{\text{Bayes}}$ is regularized even with almost flat priors. Furthermore, the VB estimator \widehat{U}^{VB} shows thresholding behavior, which leads to exact sparsity in multidimensional cases. Exact sparsity also appears in EVB learning and EBayes learning. In the subsequent sections, we explain those observations in terms of model-induced regularization and phase transitions.

7.3 Model-Induced Regularization

In this section, we explain the origin of the shrinkage of the Bayesian estimator, observed in Section 7.2. The shrinkage is caused by an implicit regularization effect, called model-induced regularization (MIR), which is strongly related to unidentifiability of statistical models.

7.3.1 Unidentifiable Models

Identifiability is formally defined as follows:

Definition 7.4 (Identifiability of statistical models) A statistical model $p(\cdot|\boldsymbol{w})$ parameterized by $\boldsymbol{w} \in \mathcal{W}$ is said to be identifiable, if the mapping $\boldsymbol{w} \mapsto p(\cdot|\boldsymbol{w})$ is one-to-one, i.e.,

$$p(\cdot|\boldsymbol{w}_1) = p(\cdot|\boldsymbol{w}_2) \iff \boldsymbol{w}_1 = \boldsymbol{w}_2 \quad \text{for any } \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}.$$

Otherwise, it is said to be unidentifiable.³

Many popular statistical models are unidentifiable.

Example 7.5 The MF model (introduced in Section 3.1) is unidentifiable, because the model distribution

$$p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{V} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2\right) \quad (7.12)$$

³ Distributions are identified in weak topology in distribution, i.e., $p(\mathbf{x}|\boldsymbol{w}_1)$ is identified with $p(\mathbf{x}|\boldsymbol{w}_2)$ if $\int f(\mathbf{x})p(\mathbf{x}|\boldsymbol{w}_1)d\mathbf{x} = \int f(\mathbf{x})p(\mathbf{x}|\boldsymbol{w}_2)d\mathbf{x}$ for any bounded continuous function $f(\mathbf{x})$.

is invariant to the following transformation $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{AT}^\top, \mathbf{BT}^{-1})$ for any nonsingular matrix $\mathbf{T} \in \mathbb{R}^{H \times H}$.

Example 7.6 The *multilayer neural network* model is unidentifiable. Consider a three-layer neural network with H hidden units:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{f}(\mathbf{x}; \mathbf{A}, \mathbf{B})\|^2\right), \quad \mathbf{f}(\mathbf{x}; \mathbf{A}, \mathbf{B}) = \sum_{h=1}^H \mathbf{b}_h \cdot \psi(\mathbf{a}_h^\top \mathbf{x}), \quad (7.13)$$

where $\mathbf{x} \in \mathbb{R}^M$ is an input vector, $\mathbf{y} \in \mathbb{R}^L$ is an output vector, $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_H) \in \mathbb{R}^{M \times H}$ and $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_H) \in \mathbb{R}^{L \times H}$ are the weight parameters to be estimated, and $\psi(\cdot)$ is an antisymmetric nonlinear activation function such as $\tanh(\cdot)$. This model expresses the identical distribution on each of the following sets of points in the parameter space:

$$\begin{aligned} & \{\mathbf{a}_h \in \mathbb{R}^M, \mathbf{b}_h = \mathbf{0}\} \cup \{\mathbf{a}_h = \mathbf{0}, \mathbf{b}_h \in \mathbb{R}^L\} \text{ for any } h, \\ & \{\mathbf{a}_h = \mathbf{a}_{h'}, \mathbf{b}_h, \mathbf{b}_{h'} \in \mathbb{R}^L, \mathbf{b}_h + \mathbf{b}_{h'} = \text{const.}\} \text{ for any pair } h, h'. \end{aligned}$$

In other words, the model is invariant for any $\mathbf{a}_h \in \mathbb{R}^M$ if $\mathbf{b}_h = \mathbf{0}$, for any $\mathbf{b}_h \in \mathbb{R}^L$ if $\mathbf{a}_h = \mathbf{0}$, and for any $\mathbf{b}_h, \mathbf{b}_{h'} \in \mathbb{R}^L$ as long as $\mathbf{b}_h + \mathbf{b}_{h'}$ is unchanged and $\mathbf{a}_h = \mathbf{a}_{h'}$.

Example 7.7 (Mixture models) The *mixture model* (introduced as Example 1.3 in Section 1.1.4) is generally unidentifiable. The model distribution is given as

$$p(\mathbf{x}|\boldsymbol{\alpha}, \{\boldsymbol{\tau}_k\}) = \sum_{k=1}^K \alpha_k p(\mathbf{x}|\boldsymbol{\tau}_k), \quad (7.14)$$

where $\mathbf{x} \in \mathcal{X}$ is an observed random variable, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\top \in \Delta^{K-1}$ and $\{\boldsymbol{\tau}_k \in \mathcal{T}\}_{k=1}^K$ are the parameters to be estimated. This model expresses the identical distribution on each of the following sets of points in the parameter space:

$$\begin{aligned} & \{\alpha_k = 0, \boldsymbol{\tau}_k \in \mathcal{T}\} \quad \text{for any } k, \\ & \{\alpha_k, \alpha_{k'} \in [0, 1], \alpha_k + \alpha_{k'} = \text{const.}, \boldsymbol{\tau}_k = \boldsymbol{\tau}_{k'}\} \text{ for any pair } k, k'. \end{aligned}$$

Namely, if the mixing weight α_k is zero for the k th mixture component, the corresponding component parameter $\boldsymbol{\tau}_k$ does not affect the model distribution, and if there are two identical components $\boldsymbol{\tau}_k = \boldsymbol{\tau}_{k'}$, the balance between the corresponding mixture weights, α_k and $\alpha_{k'}$, are arbitrary.

Readers might have noticed that, in the multilayer neural network (Example 7.6) and the mixture model (Example 7.7), the model expressed by the unidentifiable sets of points corresponds to the model with fewer components or smaller degrees of freedom. For example, if $\mathbf{a}_h = \mathbf{0}$ or $\mathbf{b}_h = \mathbf{0}$ in the neural network with H hidden units, the model is reduced to the neural network with $H - 1$ hidden units. If two hidden units receive the identical input, i.e., $\psi(\mathbf{a}_h^\top \mathbf{x}) = \psi(\mathbf{a}_{h'}^\top \mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^M$, they can be combined into a single unit with its output weight equal to the sum of the original output weights, i.e., $\mathbf{b}_h + \mathbf{b}_{h'} \rightarrow \mathbf{b}_h$. Thus, the model is again reduced to the neural network with $H - 1$ hidden units. The same applies to the mixture models and many other popular statistical models, including Bayesian networks, hidden Markov models, and latent Dirichlet allocation, which were introduced in Chapter 4. As will be explained shortly, this nesting structure—simpler models correspond to unidentifiable sets of points in the parameter space of more complex models—is essential for MIR.

7.3.2 Singularities

Continuous points denoting the same distribution are called *singularities*, on which the *Fisher information*,

$$\mathbb{S}_+^D \ni \mathbf{F} = \int \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \left(\frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \right)^\top p(\mathbf{x}|\mathbf{w}) d\mathbf{x}, \quad (7.15)$$

is *singular*, i.e., it has at least one zero eigenvalue. This is a natural consequence from the fact that the Fisher information corresponds to the *metric* when the distance between two points in the parameter space is measured by the *KL divergence* (Jeffreys, 1946), i.e., it holds that

$$\text{KL}(p(\mathbf{x}|\mathbf{w}) \| p(\mathbf{x}|\mathbf{w} + \Delta \mathbf{w})) = \frac{1}{2} \Delta \mathbf{w}^\top \mathbf{F} \Delta \mathbf{w} + O(\|\Delta \mathbf{w}\|^3)$$

for a small change $\Delta \mathbf{w}$ of the parameter. On the singularities, there is at least one direction in which the small change $\Delta \mathbf{w}$ does not affect the distribution, implying that the Fisher metric \mathbf{F} is singular. This means that the *volume element*, proportional to the determinant of the Fisher metric, is zero on the singularities, while it is positive on the regular points (see Appendix B for more details on the Fisher metric and the volume element in the parameter space).

This strong nonuniformity of (the density of) the volume element affects the behavior of Bayesian learning. For this reason, statistical models having singularities in their parameter space are called *singular models* and distinguished

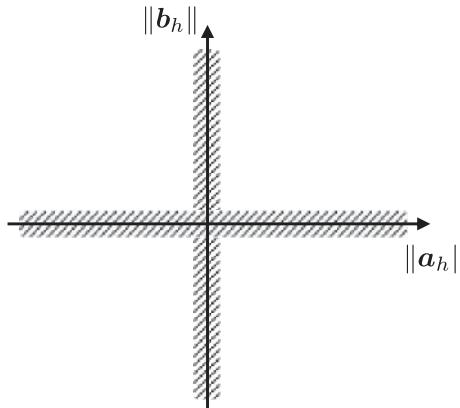


Figure 7.7 Singularities of a neural network model.

from the *regular models* in statistical learning theory (Watanabe, 2009). There are two aspects of how singularities affect the learning properties. In this chapter, we focus on one aspect that leads to MIR. The other aspect will be discussed in Chapter 13.

Figure 7.7 illustrates the singularities in the parameter space of the three-layer neural network (7.13) with $H = 1$ hidden unit (see Example 7.6). The horizontal axis corresponds to an arbitrary direction of $\mathbf{a}_h \in \mathbb{R}^M$, while the vertical axis corresponds to an arbitrary direction of $\mathbf{b}_h \in \mathbb{R}^L$. The shadowed locations correspond to the singularities. Importantly, all points on the singularities express the *identical* neural network model with no ($H = 0$) hidden unit, while each *regular* point expresses a *different* neural network model with $H = 1$ hidden unit. This illustration gives an intuition that the neighborhood of the smaller model ($H = 0$) is broader than the neighborhood of the larger model ($H = 1$) in the parameter space.

Consider the *Jeffreys prior*,

$$p^{\text{Jef}}(\mathbf{w}) \propto \sqrt{\det(\mathbf{F})}, \quad (7.16)$$

which is the *uniform prior* in the space of distributions when the distance is measured by the KL divergence (see Appendix B). As discussed previously, the Fisher information is singular on the singularities, giving $p^{\text{Jef}}(\mathbf{w}) = 0$ for the smaller model (with $H = 0$), while the Fisher information is regular on the other points, giving $p^{\text{Jef}}(\mathbf{w}) > 0$ for the larger model (with $H = 1$). Also in the neighborhood of the singularities, the Fisher information has similar values and it holds that $p^{\text{Jef}}(\mathbf{w}) \ll 1$. This means that, in comparison with the

Jeffreys prior, the flat priors on \mathbf{a}_h and \mathbf{b}_h —the uniform prior *in the parameter space*—put much more mass to the smaller model and its neighborhood. A consequence is that, if we apply Bayesian learning with the flat prior, the overweighted singularities and their neighborhood pull the estimator to the smaller model *through the integral computation*, which induces implicit regularization—MIR. The same argument holds for mixture models (Example 7.6), and other popular models, including Bayesian networks, hidden Markov models, and latent Dirichlet allocation.

In summary, MIR occurs in general singular models for the following reasons:

- There is strong nonuniformity in (the density of) the volume element around the singularities.
- Singularities correspond to the model with fewer degrees of freedom than the regular points.

This structure in the parameter space makes the flat prior favor smaller models in Bayesian learning, which appears as MIR. Note that MIR does not occur in point-estimation methods, including ML estimation and MAP learning, since the nonuniformity of the volume element affects the estimator only through *integral computations*.

MIR also occurs in the MF model (Example 7.6), which will be investigated in the next subsection with a generalization of the Jeffreys prior.

7.3.3 MIR in one-Dimensional Matrix Factorization

In Section 7.2, we numerically observed MIR—the Bayesian estimator is shrunken even with the almost flat prior in the one-dimensional MF model. However, in the MF model, the original definition (7.16) of the Jeffreys prior is zero everywhere in the parameter space because of the equivalence class structure (Figure 7.1), and therefore, it provides no information on MIR. To evaluate the nonuniformity of the volume element, we redefine the (generalized) Jeffreys prior by ignoring the zero *common* eigenvalues, i.e.,

$$p^{\text{Jef}}(\mathbf{w}) \propto \sqrt{\prod_{d=1}^{\bar{D}} \lambda_d}, \quad (7.17)$$

where λ_d is the d th largest eigenvalue of the Fisher metric \mathbf{F} , and \bar{D} is the maximum number of positive eigenvalues over the whole parameter space.

Let us consider the *nonfactorizing* model,

$$p(V|U) = \text{Gauss}_1(V; U, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(V - U)^2\right), \quad (7.18)$$

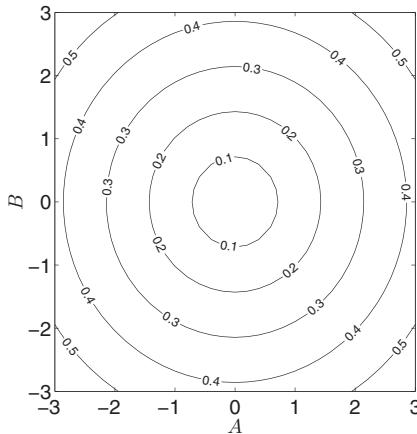


Figure 7.8 The (unnormalized) Jeffreys noninformative prior (7.20) of the one-dimensional MF model (7.5).

where U itself is the parameter to be estimated. The Jeffreys prior for this model is uniform (see Example B.1 in Appendix B for derivation):

$$p^{\text{Jef}}(U) \propto 1. \quad (7.19)$$

On the other hand, the Jeffreys prior for the MF model (7.5) is given as follows (see Example B.2 in Appendix B for derivation):

$$p^{\text{Jef}}(A, B) \propto \sqrt{A^2 + B^2}, \quad (7.20)$$

which is illustrated in Figure 7.8. Note that the Jeffreys priors (7.19) and (7.20) for both cases are *improper*, meaning that they cannot be normalized since their integrals diverge.

Jeffreys (1946) stated that the both combinations, the *nonfactorizing* model (7.18) with its Jeffreys prior (7.19) and the MF model (7.5) with its Jeffreys prior (7.20) give the equivalent Bayesian estimator. We can easily show that the former combination, Eqs. (7.18) and (7.19), gives an unregularized solution. Thus, the Bayesian estimator in the MF model (7.5) with its Jeffreys prior (7.20) is also unregularized. Since the flat prior on (A, B) has more probability mass around the origin than the Jeffreys prior (7.20) (see Figure 7.8), it favors smaller $|U|$ and regularizes the Bayesian estimator.

Although MIR appears also in regular models unless the Jeffreys prior is flat in the parameter space, its effect is prominent in singular models with unidentifiability, since the difference between the flat prior and the Jeffreys prior is large.

7.3.4 Evidence View of Unidentifiable Models

MIR works as Occam's razor in general. MacKay (1992) explained, with the illustration shown in the left panel of Figure 7.9, that evidence-based (i.e., free-energy-minimization-based) model selection is naturally equipped with Occam's razor. In the figure, the horizontal axis denotes the space of the observed data set \mathcal{D} . \mathcal{H}_1 and \mathcal{H}_2 denote a simple hypothesis and a more complex hypothesis, respectively. For example, in the MF model, the observed data set corresponds to the observed matrix, i.e., $\mathcal{D} = V$, \mathcal{H}_1 corresponds to a lower-rank model, and \mathcal{H}_2 corresponds to a higher-rank model. The vertical axis indicates the evidence or marginal likelihood,

$$p(\mathcal{D}|\mathcal{H}_t) = \langle p(\mathcal{D}|\boldsymbol{w}_{\mathcal{H}_t}, \mathcal{H}_t) \rangle_{p(\boldsymbol{w}_{\mathcal{H}_t})} \quad \text{for } t = 1, 2, \quad (7.21)$$

where $\boldsymbol{\theta}_{\mathcal{H}_t}$ denotes the unknown parameters that the hypothesis \mathcal{H}_t has.

Since \mathcal{H}_1 is simple, it covers a limited area of the space of \mathcal{D} (meaning that it can explain only a simple phenomenon), while \mathcal{H}_2 covers a broader area. The illustration implies that, because of the normalization, it holds that

$$p(\mathcal{D}|\mathcal{H}_1) > p(\mathcal{D}|\mathcal{H}_2) \quad \text{for } \mathcal{D} \in C_1,$$

where C_1 denotes the observed data region where \mathcal{H}_1 can explain the data well. This view gives an intuition on why evidence-based model selection prefers simpler models when the observed data can be well explained by them.

However, this view does not explain MIR, which is observed even without any model selection procedure. In fact, the illustration in the left panel of Figure 7.9 is not accurate for unidentifiable models unless the Jeffreys prior is adopted (note that a hypothesis consists of a model and a prior). The right illustration of Figure 7.9 is a more accurate view for unidentifiable models. When \mathcal{H}_2 is a complex unidentifiable model nesting \mathcal{H}_1 as a simpler model in

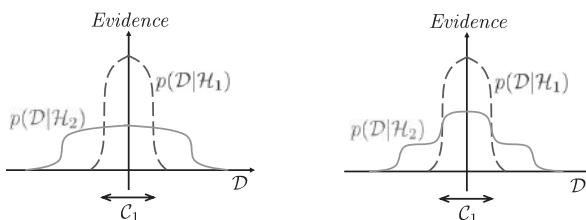


Figure 7.9 Left: The evidence view by MacKay (1992), which gives an intuition on why evidence-based model selection prefers simpler models. Right: A more accurate view for *unidentifiable* models. Simpler models are preferred *even without explicit model selection*.

its parameter space, its evidence $p(\mathcal{D}|\mathcal{H}_2)$ has a bump covering the region C_1 if the flat prior is adopted. This is because the flat prior typically places large weights on the singularities representing the simpler model \mathcal{H}_1 .

7.4 Phase Transition in VB Learning

In Section 7.3, we explained MIR, which shrinks the Bayesian estimator. We can expect that VB learning, which involves integral computations, inherits this property. However, we observe in Figure 7.6 that VB learning behaves differently from Bayesian learning. Actually, the Bayesian estimator behaves more similarly to the ML estimator (the MAP estimator with almost flat priors), rather than the VB estimator. A remarkable difference is that the VB estimator, which is upper-bounded by the PJS estimator (7.2), shows exact sparsity, i.e., the estimator can be zero for nonzero observation $|V|$. In this section, we explain that this gap is caused by a phase transition phenomenon in VB learning.

The middle graph in Figure 7.2 shows the Bayes posterior when $V = 1$. The probability mass in the first and the third quadrants pulls the product $U = BA$ toward the positive direction, and the mass in the second and the fourth quadrants toward the negative direction. Since the Bayes posterior is skewed and more mass is placed in the first and the third quadrants, the Bayesian estimator $\widehat{U}^{\text{Bayes}} = \langle BA \rangle_{p(A,B|V)}$ is positive. This is true even if $V > 0$ is very small, and therefore, no thresholding occurs in Bayesian learning—the Bayesian estimator is not sparse.

On the other hand, the VB posterior (the top-right graph of Figure 7.4) is prohibited to be skewed because of the independent constraint, which causes the following phase transition phenomenon. As seen in Figure 7.2, the Bayes posterior has two modes unless $V = 0$, and the distance between the two modes increases as $|V|$ increases. Since the VB posterior tries to approximate the Bayes posterior with a single uncorrelated distribution, it stays at the origin if the two modes are close to each other so that covering both modes minimizes the free energy. The VB posterior detaches from the origin if the two modes get far apart so that approximating either one of the modes minimizes the free energy. This phase transition mechanism makes the VB estimator exactly sparse. The profile of the Bayes posterior (the middle graph of Figure 7.2) implies that, if we restrict the posterior to be Gaussian, but allow it to have correlation between A and B , exact sparsity will not appear. In this sense, we can say that MIR is enhanced by the independence constraint, which was imposed for computational tractability.

Mackay (2001) pointed out that there are cases where VB learning prunes model components *inappropriately*, by giving a toy example of a mixture of Gaussians. Note that *appropriateness* was measured in terms of the similarity to full Bayesian learning. He plotted the free energy of the mixture of Gaussians as a function of hidden responsibility variables—the probabilities that each sample belongs to each Gaussian component—and argued that VB learning sometimes favors simpler models too much. In this case, degrees of freedom are pruned when spontaneous symmetry breaking (a phase transition) occurs. Interestingly, in the MF model, degrees of freedom are pruned when spontaneous symmetry breaking does *not* occur, as explained earlier.

Eq. (7.3) implies that the symmetry breaking occurs when $V > \underline{\gamma}_h^{\text{VB}} \approx \sqrt{M\sigma^2} = 1$, which coincides with the average contribution of noise to the observed singular values over all singular components—more accurately, $\sqrt{M\sigma^2}$ is the square root of the average eigenvalues of the Wishart matrix $\mathcal{E}\mathcal{E}^\top \sim \text{Wishart}_L(\sigma^2\mathbf{I}_L, M)$.⁴ In this way, VB learning discards singular components dominated by noise.

Given that the full Bayesian estimator in MF is not sparse (see Figure 7.6), one might argue that the sparsity of VB learning is an *inappropriate* artifact. On the other hand, given that automatic model pruning by VB learning has been acknowledged as a practically useful property (Bishop, 1999b; Bishop and Tipping, 2000; Sato et al., 2004; Babacan et al., 2012b), one might also argue that *appropriateness* should be measured in terms of performance. Motivated by the latter idea, performance analysis has been carried out (Nakajima et al., 2015), which will be detailed in Chapter 8.

In the empirical Bayesian scenario, where the prior variances c_a, c_b are also estimated from observation, Bayesian learning also gives a sparse solution, which is shown as diamonds (labeled as “EBayes”) in Figure 7.6. This is somewhat natural since, in empirical Bayesian learning, the dependency between A and c_a^{-2} (as well as B and c_b^{-2}) in the prior (7.6) (in the prior (7.7)) and hence in the Bayes posterior is broken—the point-estimation of c_a^2 (as well as c_b^2) forces it to be independent of all other parameters. This forced independence causes a similar phase transition phenomenon to the one caused by the independence constraint between A and B in the (nonempirical) VB learning, and results in exact sparsity of the EBayes estimator.

EVB learning has a different transition point, and tends to give a sparser solution than VB learning. A notable difference from the VB estimator is that the EVB estimator is no longer continuous as a function of the observation V . This comes from the fact that, when $|V| > \underline{\gamma}_{\text{local-EVB}}$, there exist two local

⁴ It holds that $\mathcal{E}\mathcal{E}^\top \sim \text{Wishart}_L(\sigma^2\mathbf{I}_L, M)$ if $\mathcal{E} \sim \text{Gauss}_L(\mathbf{0}, \sigma^2\mathbf{I}_L)$.

solutions (see Figure 6.3), but the global solution is $\widehat{U}^{\text{EVB}} = 0$ until the observed amplitude $|V|$ exceeds $\underline{\gamma}^{\text{EVB}} (> \underline{\gamma}^{\text{local-EVB}})$. When the positive local solution $\check{\gamma}^{\text{EVB}}$ becomes the global solution, it is already distant from the origin, which makes the estimator noncontinuous at the thresholding point (see the dashed curve labeled as “EVB” in Figure 7.6).

7.5 Factorization as ARD Model

As shown in Section 7.1, MIR in VB learning for the MF model appears as PJS shrinkage. We can see this as a natural consequence from the equivalence between the MF model and the ARD model (Neal, 1996).

Assume that $C_A = I_H$ in the MF model (6.1) through (6.3), and consider the following transformation: $\mathbf{B}\mathbf{A}^\top \mapsto \mathbf{U} \in \mathbb{R}^{L \times M}$. Then, the likelihood (6.1) and the prior (6.2) on \mathbf{A} become

$$p(\mathbf{V}|\mathbf{U}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{V} - \mathbf{U}\|_{\text{Fro}}^2\right), \quad (7.22)$$

$$p(\mathbf{U}|\mathbf{B}) \propto \exp\left(-\frac{1}{2} \text{tr}\left(\mathbf{U}^\top (\mathbf{B}\mathbf{B}^\top)^\dagger \mathbf{U}\right)\right), \quad (7.23)$$

where \dagger denotes the Moore–Penrose generalized inverse of a matrix. The prior (6.3) on \mathbf{B} is kept unchanged. $p(\mathbf{U}|\mathbf{B})$ in Eq. (7.23) is so-called the ARD prior with the covariance hyperparameter $\mathbf{B}\mathbf{B}^\top \in \mathbb{R}^{L \times L}$. It is known that this prior induces the ARD property—empirical Bayesian learning, where the prior covariance hyperparameter $\mathbf{B}\mathbf{B}^\top$ is estimated from observation by maximizing the marginal likelihood (or minimizing the free energy), induces strong regularization and sparsity (Neal, 1996). Efron and Morris (1973) showed that this particular model gives the JS shrinkage estimator as an empirical Bayesian estimator (see Appendix A).

This equivalence can explain the sparsity-inducing terms (3.113) through (3.116), introduced for sparse additive matrix factorization (SAMF) in Section 3.5. The ARD prior (7.23) induces low-rankness on \mathbf{U} if no restriction on $\mathbf{B}\mathbf{B}^\top$ is imposed. We can similarly show that, $(\gamma_l^e)^2$ in Eq. (3.114) corresponds to the prior variance shared by the entries in $\tilde{\mathbf{u}}_l \equiv \gamma_l^e \tilde{\mathbf{d}}_l \in \mathbb{R}^M$, that $(\gamma_m^d)^2$ in Eq. (3.115) corresponds to the prior variance shared by the entries in $\mathbf{u}_m \equiv \gamma_m^d \mathbf{e}_m \in \mathbb{R}^L$, and that $E_{l,m}^2$ in Eq. (3.116) corresponds to the prior variance on $U_{l,m} \equiv E_{l,m} D_{l,m} \in \mathbb{R}$, respectively. This explains why the factorization forms in Eqs. (3.113) through (3.116) induce low-rank, rowwise, columnwise, and elementwise sparsity, respectively. If we employ the sparse matrix factorization (SMF) term (3.117), ARD occurs in each partition, which induces partitionwise sparsity and low-rank sparsity within each partition.

8

Performance Analysis of VB Matrix Factorization

In this chapter, we further analyze the behavior of VB learning in the fully observed MF model, introduced in Section 3.1. Then, we derive a theoretical guarantee for rank estimation (Nakajima et al., 2015), which corresponds to the hidden dimensionality selection in principal component analysis (PCA).

In Chapter 6, we derived an analytic-form solution (Theorem 6.13) of EVB learning, where the prior variances are also estimated from observation. When discussing the dimensionality selection performance in PCA, it is more practical to assume that the noise variance σ^2 is estimated, since it is unknown in many situations. To this end, we first analyze the behavior of the noise variance estimator. After that, based on the random matrix theory, we derive a theoretical guarantee of dimensionality selection performance, and show numerical results validating the theory. We also discuss the relation to an alternative dimensionality selection method (Hoyle, 2008) based on the Laplace approximation.

In the following analysis, we use some results in Chapter 6. Specifically, we mostly rely on Theorem 6.13 along with the corollaries and the equations summarized in Section 6.10.

8.1 Objective Function for Noise Variance Estimation

Let us consider the *complete* empirical VB problem, where all the variational parameters and the hyperparameters are estimated in the free energy minimization framework:

$$\begin{aligned} & \min_{\{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2, c_{a_h}^2, c_{b_h}^2\}_{h=1}^H, \sigma^2} F \\ \text{s.t. } & \{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \quad \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2, c_{a_h}^2, c_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H, \sigma^2 \in \mathbb{R}_{++}. \end{aligned} \tag{8.1}$$

Here, the free energy is given by

$$2F = LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^L \gamma_h^2}{\sigma^2} + \sum_{h=1}^H 2F_h, \quad (8.2)$$

$$\text{where } 2F_h = M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2}{c_{b_h}^2} - (L+M) + \frac{-2\widehat{a}_h\widehat{b}_h\gamma_h + (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2)(\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2)}{\sigma^2}. \quad (8.3)$$

Note that we are focusing on the solution with diagonal posterior covariances without loss of generality (see Theorem 6.4).

We have already obtained the empirical VB estimator (Theorem 6.13) and the minimum free energy (Corollary 6.21) for given σ^2 . By using those results, we can express the free energy (8.2) as a function of σ^2 . With the rescaled expressions (6.132) through (6.138), the free energy can be written in a simple form, which leads to the following theorem:

Theorem 8.1 *The noise variance estimator, denoted by $\widehat{\sigma}^{2 \text{ EVB}}$, is the global minimizer of*

$$\Omega(\sigma^{-2}) \left(\equiv \frac{2F(\sigma^{-2})}{LM} + \text{const.} \right) = \frac{1}{L} \left(\sum_{h=1}^H \psi \left(\frac{\gamma_h^2}{M\sigma^2} \right) + \sum_{h=H+1}^L \psi_0 \left(\frac{\gamma_h^2}{M\sigma^2} \right) \right), \quad (8.4)$$

where

$$\psi(x) = \psi_0(x) + \theta(x > \underline{x}) \psi_1(x), \quad (8.5)$$

$$\psi_0(x) = x - \log x, \quad (8.6)$$

$$\psi_1(x) = \log(\tau(x; \alpha) + 1) + \alpha \log \left(\frac{\tau(x; \alpha)}{\alpha} + 1 \right) - \tau(x; \alpha). \quad (8.7)$$

Here, \underline{x} is given by

$$\underline{x} = \left(1 + \frac{\alpha}{\tau} \right) \left(1 + \frac{\alpha}{\tau} \right), \quad (8.8)$$

where $\underline{\tau}$ is defined in Theorem 6.13, $\tau(x; \alpha)$ is a function of $x (> \underline{x})$ defined by

$$\tau(x; \alpha) = \frac{1}{2} \left(x - (1 + \alpha) + \sqrt{(x - (1 + \alpha))^2 - 4\alpha} \right), \quad (8.9)$$

and $\theta(\cdot)$ denotes the indicator function such that $\theta(\text{condition}) = 1$ if the condition is true and $\theta(\text{condition}) = 0$ otherwise.

Proof By using Lemma 6.14 and Lemma 6.16, the free energy (8.2) can be written as a function of σ^2 as follows:

$$2F = LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^L \gamma_h^2}{\sigma^2} + \sum_{h=1}^H \theta(\gamma_h > \underline{\gamma}^{\text{EVB}}) F_h^{\text{EVB-Posi}}, \quad (8.10)$$

where $F_h^{\text{EVB-Posi}}$ is given by Eq. (6.112). By using Eqs. (6.133) and (6.135), Eq. (6.112) can be written as

$$\begin{aligned} F_h^{\text{EVB-Posi}} &= M \log(\tau_h + 1) + L \log\left(\frac{\tau_h}{\alpha} + 1\right) - M\tau_h \\ &= M\psi_1(x_h). \end{aligned} \quad (8.11)$$

Therefore, Eq. (8.10) is written as

$$\begin{aligned} 2F &= M \left\{ \sum_{h=1}^L \log\left(\frac{2\pi\gamma_h^2}{M}\right) + \sum_{h=1}^L \left(\log\left(\frac{M\sigma^2}{\gamma_h^2}\right) + \frac{\gamma_h^2}{M\sigma^2} \right) \right. \\ &\quad \left. + \sum_{h=1}^H \theta(\gamma_h > \underline{\gamma}^{\text{EVB}}) \frac{F_h^{\text{EVB-Posi}}}{M} \right\} \\ &= M \left\{ \sum_{h=1}^L \log\left(\frac{2\pi\gamma_h^2}{M}\right) + \sum_{h=1}^L \psi_0(x_h) + \sum_{h=1}^H \theta(x_h > \underline{x}) \psi_1(x_h) \right\}. \end{aligned}$$

Note that the first term in the curly braces is constant with respect to σ^2 . By defining

$$\mathcal{Q} = \frac{2F}{LM} - \frac{1}{L} \sum_{h=1}^L \log\left(\frac{2\pi\gamma_h^2}{M}\right),$$

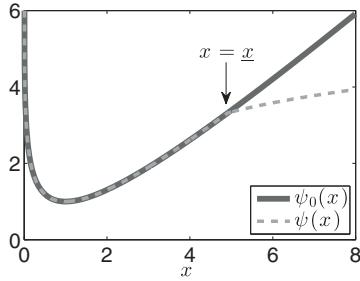
we obtain Eq. (8.4), which completes the proof of Theorem 8.1. \square

The functions $\psi_0(x)$ and $\psi(x)$ are depicted in Figure 8.1. We can confirm the convexity of $\psi_0(x)$ and the quasiconvexity of $\psi(x)$ (Lemma 8.4 in Section 8.3), which are useful properties in the subsequent analysis.¹

8.2 Bounds of Noise Variance Estimator

Let \widehat{H}^{EVB} be the estimated rank by EVB learning, i.e., the rank of the EVB estimator \widehat{U}^{EVB} , such that $\widehat{\gamma}_h^{\text{EVB}} > 0$ for $h = 1, \dots, \widehat{H}^{\text{EVB}}$, and $\widehat{\gamma}_h^{\text{EVB}} = 0$ for

¹ A function $f : \mathcal{X} \mapsto \mathbb{R}$ on the domain \mathcal{X} being a convex subset of a real vector space is said to be *quasiconvex* if $f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \max(f(\mathbf{x}_1), f(\mathbf{x}_2))$ for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$. It is furthermore said to be *strictly quasiconvex* if $f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) < \max(f(\mathbf{x}_1), f(\mathbf{x}_2))$ for all $\mathbf{x}_1 \neq \mathbf{x}_2$ and $\lambda \in (0, 1)$. Intuitively, a strictly quasiconvex function does not have more than one local minima.

Figure 8.1 $\psi_0(x)$ and $\psi(x)$.

$h = \widehat{H}^{\text{EVB}} + 1, \dots, H$. By further analyzing the objective (8.4), we can derive bounds of the estimated rank and the noise variance estimator:

Theorem 8.2 \widehat{H}^{EVB} is upper-bounded as

$$\widehat{H}^{\text{EVB}} \leq \overline{H} = \min\left(\left\lceil \frac{L}{1+\alpha} \right\rceil - 1, H\right), \quad (8.12)$$

and the noise variance estimator $\widehat{\sigma}^2^{\text{EVB}}$ is bounded as follows:

$$\max\left(\underline{\sigma}_{\overline{H}+1}^2, \frac{\sum_{h=\overline{H}+1}^L \gamma_h^2}{M(L-\overline{H})}\right) \leq \widehat{\sigma}^2^{\text{EVB}} \leq \frac{1}{LM} \sum_{h=1}^L \gamma_h^2, \quad (8.13)$$

$$\text{where } \underline{\sigma}_h^2 = \begin{cases} \infty & \text{for } h = 0, \\ \frac{\gamma_h^2}{M_x} & \text{for } h = 1, \dots, L, \\ 0 & \text{for } h = L+1. \end{cases} \quad (8.14)$$

Theorem 8.2 states that EVB learning discards the $(L - \lceil L/(1+\alpha) \rceil + 1)$ smallest components, regardless of the observed singular values $\{\gamma_h\}_{h=1}^L$. For example, half of the components are always discarded when the matrix is square (i.e., $\alpha = L/M = 1$). The smallest singular value γ_L is always discarded, and $\widehat{\sigma}^2^{\text{EVB}} \geq \gamma_L^2/M$ always holds.

Given the EVB estimators $\{\widehat{\gamma}_h^{\text{EVB}}\}_{h=1}^H$ for the singular values, the noise variance estimator $\widehat{\sigma}^2^{\text{EVB}}$ is specified by the following corollary:

Corollary 8.3 The EVB estimator for the noise variance satisfies the following equality:

$$\widehat{\sigma}^2^{\text{EVB}} = \frac{1}{LM} \left(\sum_{l=1}^L \gamma_l^2 - \sum_{h=1}^H \gamma_h \widehat{\gamma}_h^{\text{EVB}} \right). \quad (8.15)$$

This corollary can be used for implementing a global EVB solver (see Chapter 9). In the next section we give the proofs of the theorem and the corollary.

8.3 Proofs of Theorem 8.2 and Corollary 8.3

First, we show nice properties of the functions, $\psi(x)$ and $\psi_0(x)$, which are defined by Eqs. (8.5) and (8.6), respectively, and depicted in Figure 8.1:

Lemma 8.4 *The following hold for $x > 0$: $\psi_0(x)$ is differentiable and strictly convex; $\psi(x)$ is continuous and strictly quasiconvex; $\psi(x)$ is differentiable except $x = \underline{x}$, at which $\psi(x)$ has a discontinuously decreasing derivative, i.e., $\lim_{x \rightarrow \underline{x}-0} \partial\psi/\partial x > \lim_{x \rightarrow \underline{x}+0} \partial\psi/\partial x$; both of $\psi_0(x)$ and $\psi(x)$ are minimized at $x = 1$. For $x > \underline{x}$, $\psi_1(x)$ is negative and decreasing.*

Proof Since

$$\begin{aligned}\frac{\partial\psi_0}{\partial x} &= 1 - \frac{1}{x}, \\ \frac{\partial^2\psi_0}{\partial x^2} &= \frac{1}{x^2} > 0,\end{aligned}\tag{8.16}$$

$\psi_0(x)$ is differentiable and strictly convex for $x > 0$ with its minimizer at $x = 1$. $\psi_1(x)$ is continuous for $x \geq \underline{x}$, and Eq. (8.11) implies that $\psi_1(x_h) \propto F_h^{\text{EVB-Posi}}$. Accordingly, $\psi_1(x) \leq 0$ for $x \geq \underline{x}$, where the equality holds when $x = \underline{x}$. This equality implies that $\psi(x)$ is continuous. Since $\underline{x} > 1$, $\psi(x)$ shares the same minimizer as $\psi_0(x)$ at $x = 1$ (see Figure 8.1).

Hereafter, we investigate $\psi_1(x)$ and $\psi(x)$ for $x \geq \underline{x}$. By differentiating Eqs. (8.7) and (6.135), respectively, we have

$$\frac{\partial\psi_1}{\partial\tau} = -\left(\frac{\frac{\tau^2}{\alpha} - 1}{(\tau + 1)\left(\frac{\tau}{\alpha} + 1\right)}\right) < 0,\tag{8.17}$$

$$\frac{\partial\tau}{\partial x} = \frac{1}{2}\left(1 + \frac{x - (1 + \alpha)}{\sqrt{(x - (1 + \alpha))^2 - 4\alpha}}\right) > 0.\tag{8.18}$$

Substituting Eq. (6.134) into Eq. (8.18), we have

$$\frac{\partial\tau}{\partial x} = \frac{\tau^2}{\alpha\left(\frac{\tau^2}{\alpha} - 1\right)}.\tag{8.19}$$

Multiplying Eqs. (8.17) and (8.19) gives

$$\frac{\partial\psi_1}{\partial x} = \frac{\partial\psi_1}{\partial\tau} \frac{\partial\tau}{\partial x} = -\left(\frac{\tau^2}{\alpha(\tau+1)(\frac{\tau}{\alpha}+1)}\right) = -\frac{\tau}{x} < 0, \quad (8.20)$$

which implies that $\psi_1(x)$ is decreasing for $x > \underline{x}$.

Let us focus on the thresholding point of $\psi(x)$ at $x = \underline{x}$. Eq. (8.20) does not converge to zero for $x \rightarrow \underline{x} + 0$ but stay negative. On the other hand, $\psi_0(x)$ is differentiable at $x = \underline{x}$. Consequently, $\psi(x)$ has a discontinuously decreasing derivative, i.e., $\lim_{x \rightarrow \underline{x}-0} \partial\psi/\partial x > \lim_{x \rightarrow \underline{x}+0} \partial\psi/\partial x$, at $x = \underline{x}$.

Finally, we prove the strict quasiconvexity of $\psi(x)$. Taking the sum of Eqs. (8.16) and (8.20) gives

$$\frac{\partial\psi}{\partial x} = \frac{\partial\psi_0}{\partial x} + \frac{\partial\psi_1}{\partial x} = 1 - \frac{1+\tau}{x} = 1 - \frac{1+\tau}{1+\tau+\alpha+\alpha\tau^{-1}} > 0.$$

This means that $\psi(x)$ is increasing for $x > \underline{x}$. Since $\psi_0(x)$ is strictly convex and increasing at $x = \underline{x}$, and $\psi(x)$ is continuous, $\psi(x)$ is strictly quasiconvex. This completes the proof of Lemma 8.4. \square

Lemma 8.4 implies that our objective (8.4) is a sum of quasiconvex functions with respect to σ^{-2} . Therefore, its minimizer can be bounded by the smallest one and the largest one among the set collecting the minimizer from each quasiconvex function:

Lemma 8.5 $\mathcal{Q}(\sigma^{-2})$ has at least one global minimizer, and any of its local minimizers is bounded as

$$\frac{M}{\gamma_1^2} \leq \widehat{\sigma}^{-2} \leq \frac{M}{\gamma_L^2}. \quad (8.21)$$

Proof The strict convexity of $\psi_0(x)$ and the strict quasiconvexity of $\psi(x)$ also hold for $\psi_0(\gamma_h^2\sigma^{-2}/M)$ and $\psi(\gamma_h^2\sigma^{-2}/M)$ as functions of σ^{-2} (for $\gamma_h > 0$). Because of the different scale factor γ_h^2/M for each $h = 1, \dots, L$, each of $\psi_0(\gamma_h^2\sigma^{-2}/M)$ and $\psi(\gamma_h^2\sigma^{-2}/M)$ has a minimizer at a different position:

$$\sigma^{-2} = \frac{M}{\gamma_h^2}.$$

The strict quasiconvexity of ψ_0 and ψ guarantees that $\mathcal{Q}(\sigma^{-2})$ is decreasing for

$$0 < \sigma^{-2} < \frac{M}{\gamma_1^2}, \quad (8.22)$$

and increasing for

$$\frac{M}{\gamma_L^2} < \sigma^{-2} < \infty. \quad (8.23)$$

This proves Lemma 8.5. \square

$\Omega(\sigma^{-2})$ has at most H nondifferentiable points, which come from the nondifferentiable point $x = \underline{x}$ of $\psi(x)$. The values

$$\underline{\sigma}_h^{-2} = \begin{cases} 0 & \text{for } h = 0, \\ \frac{M_x}{\gamma_h^2} & \text{for } h = 1, \dots, L, \\ \infty & \text{for } h = L + 1, \end{cases} \quad (8.24)$$

defined in Eq. (8.14) for $h = 1, \dots, H$ actually correspond to those points.

Lemma 8.4 states that, at $x = \underline{x}$, $\psi(x)$ has a discontinuously decreasing derivative and neither $\psi_0(x)$ nor $\psi(x)$ has a discontinuously increasing derivative at any point. Therefore, none of those nondifferentiable points can be a local minimum. Consequently, we have the following lemma:

Lemma 8.6 $\Omega(\sigma^{-2})$ has no local minimizer at $\sigma^{-2} = \underline{\sigma}_h^{-2}$ for $h = 1, \dots, H$, and therefore any of its local minimizers is a stationary point.

Then, Theorem 6.13 leads to the following lemma:

Lemma 8.7 The estimated rank is $\widehat{H} = h$ if and only if the inverse noise variance estimator lies in the range

$$\widehat{\sigma}^{-2} \in \mathcal{B}_h \equiv \left\{ \sigma^{-2}; \underline{\sigma}_h^{-2} < \sigma^{-2} < \underline{\sigma}_{h+1}^{-2} \right\}. \quad (8.25)$$

Figure 8.2 shows quasiconvex functions $\{\psi(\gamma_h^2 \sigma^{-2}/M)\}_{h=1}^H$ and their average $\Omega(\sigma^{-2})$ in two exemplary cases for $H = L$. In the left case, the inverse noise variance estimator $\widehat{\sigma}^{-2}$ is smaller than the inverse threshold $\underline{\sigma}_1^{-2}$ for the largest singular value, and therefore no EVB estimator $\widehat{\gamma}_h$ is positive, i.e., $\widehat{H} = 0$. In the right case, it holds that $\underline{\sigma}_1^{-2} < \widehat{\sigma}^{-2} < \underline{\sigma}_2^{-2}$, and therefore $\widehat{\gamma}_1$ is positive and the others are zero, i.e., $\widehat{H} = 1$.

We have the following lemma:

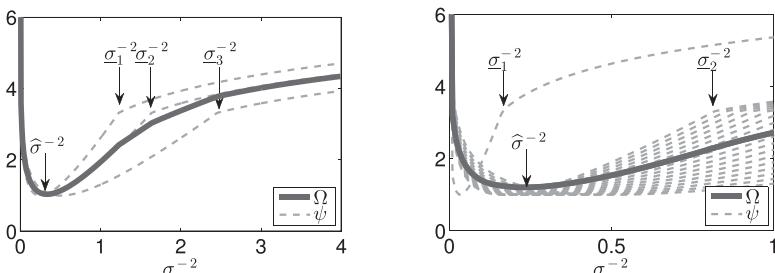


Figure 8.2 $\{\psi(\gamma_h^2 \sigma^{-2}/M)\}_{h=1}^H$ and $\Omega(\sigma^{-2})$ in two exemplary cases for $H = L$. Left: the case where $\gamma_h^2/M = 4, 3, 2$ for $h = 1, 2, 3$. Right: the case where $\gamma_1^2/M = 30$, $\gamma_h^2/M = 6, 5.75, 5.5, \dots, 2.0$ for $h = 2, \dots, 18$.

Lemma 8.8 *The derivative of $\Omega(\sigma^{-2})$ is given by*

$$\Theta \equiv \frac{\partial \Omega}{\partial \sigma^{-2}} = -\sigma^2 + \frac{\sum_{h=1}^{\widehat{H}} \gamma_h (\gamma_h - \check{\gamma}_h^{\text{EVB}}) + \sum_{h=\widehat{H}+1}^L \gamma_h^2}{LM}, \quad (8.26)$$

where \widehat{H} is a function of σ^{-2} defined by

$$\widehat{H} = \widehat{H}(\sigma^{-2}) = h \quad \text{if} \quad \sigma^{-2} \in \mathcal{B}_h. \quad (8.27)$$

Proof The derivative of Eq. (8.4) with respect to σ^{-2} is given by

$$\frac{\partial \Omega}{\partial \sigma^{-2}} = \frac{1}{L} \left(\sum_{h=1}^H \frac{\gamma_h^2}{M} \frac{\partial \psi}{\partial x} + \sum_{h=H+1}^L \frac{\gamma_h^2}{M} \frac{\partial \psi_0}{\partial x} \right). \quad (8.28)$$

By using Eqs. (8.16) and (8.20), Eq. (8.28) can be written as

$$\begin{aligned} \frac{\partial \Omega}{\partial \sigma^{-2}} &= \frac{1}{L} \left(\sum_{h=1}^L \frac{\gamma_h^2}{M} \frac{\partial \psi_0}{\partial x} + \sum_{h=1}^H \theta(x_h \geq \underline{x}) \frac{\gamma_h^2}{M} \frac{\partial \psi_1}{\partial x} \right) \\ &= \frac{1}{L} \left(\sum_{h=1}^L \frac{\gamma_h^2}{M} \left(1 - \frac{1}{x_h} \right) - \sum_{h=1}^H \theta(x_h \geq \underline{x}) \frac{\gamma_h^2 \tau_h}{M x_h} \right) \\ &= \frac{\sum_{h=1}^L \gamma_h^2}{LM} - \sigma^2 - \frac{1}{L} \sum_{h=1}^H \theta(\tau_h \geq \tau) \sigma^2 \tau_h. \end{aligned} \quad (8.29)$$

Here we also used the definition (6.132) of x_h . Using Eq. (6.133), Eq. (8.29) can be written as

$$\begin{aligned} \frac{\partial \Omega}{\partial \sigma^{-2}} &= \frac{\sum_{h=1}^L \gamma_h^2}{LM} - \sigma^2 - \sum_{h=1}^H \theta(\gamma_h \geq \underline{\gamma}_h^{\text{EVB}}) \frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{LM} \\ &= -\sigma^2 + \frac{\sum_{h=1}^H \gamma_h (\gamma_h - \check{\gamma}_h^{\text{EVB}}) + \sum_{h=H+1}^L \gamma_h^2}{LM}. \end{aligned} \quad (8.30)$$

Here we also used the definition (6.101) of $\check{\gamma}_h^{\text{EVB}}$. Using the definition (8.27) and Lemma 8.7, we can replace $\check{\gamma}_h^{\text{EVB}}$ and H with $\check{\gamma}_h^{\text{EVB}}$ and \widehat{H} , respectively, which completes the proof of Lemma 8.8. \square

Note that Eq. (8.26) involves the shrinkage estimator $\check{\gamma}_h^{\text{EVB}}$, which is a function of σ^{-2} (see Eq. (6.103)). For each hypothetical \widehat{H} , the solutions of the equation

$$\Theta = 0 \quad (8.31)$$

lying in $\sigma^{-2} \in \mathcal{B}_{\widehat{H}}$ are stationary points, and hence candidates for the global minimum. If we can solve Eq. (8.31) for all $\widehat{H} = 1, \dots, H$, we can obtain the global solution by evaluating the objective (8.4) at each obtained

stationary point. However, solving Eq. (8.31) is computationally hard unless \widehat{H} is small.² Based on Lemma 8.8, we will obtain tighter bounds than Lemma 8.5.

Since

$$\gamma_h - \check{\gamma}_h^{\text{EVB}} > 0,$$

Eq. (8.26) is upper-bounded by

$$\Theta \leq -\sigma^2 + \sum_{h=1}^L \frac{\gamma_h^2}{LM},$$

which leads to the upper-bound given in Eq. (8.13). Actually, if

$$\left(\sum_{h=1}^L \frac{\gamma_h^2}{LM} \right)^{-1} \in \mathcal{B}_0,$$

then

$$\begin{aligned} \widehat{H} &= 0, \\ \widehat{\sigma}^2 &= \sum_{h=1}^L \frac{\gamma_h^2}{LM}, \end{aligned}$$

is a local minimum.

The following lemma is easily obtained from Eq. (6.103) by using the inequalities $z_1 < \sqrt{z_1^2 - z_2^2} < z_1 - z_2$ for $z_1 > z_2 > 0$:

Lemma 8.9 *For $\gamma_h \geq \underline{\gamma}^{\text{EVB}}$, the EVB shrinkage estimator (6.103) can be bounded as follows:*

$$\gamma_h - \frac{(\sqrt{M} + \sqrt{L})^2 \sigma^2}{\gamma_h} < \check{\gamma}_h^{\text{EVB}} < \gamma_h - \frac{(M + L)\sigma^2}{\gamma_h}. \quad (8.32)$$

This lemma is important for our analysis, because it allows us to bound the most complicated part of Eq. (8.26) by quantities independent of γ_h , i.e.,

$$(M + L)\sigma^2 < \gamma_h \left(\gamma_h - \check{\gamma}_h^{\text{EVB}} \right) < (\sqrt{M} + \sqrt{L})^2 \sigma^2. \quad (8.33)$$

Using Eq. (8.33), we obtain the following lemma:

Lemma 8.10 *Any local minimizer exists in $\sigma^{-2} \in \mathcal{B}_{\widehat{H}}$ such that*

$$\widehat{H} < \frac{L}{1 + \alpha},$$

² It is easy to derive a closed-form solution for $\widehat{H} = 0, 1$.

and the following holds for any local minimizer lying in $\sigma^{-2} \in \mathcal{B}_{\widehat{H}}$:

$$\widehat{\sigma}^2 \geq \frac{\sum_{h=\widehat{H}+1}^L \gamma_h^2}{LM - \widehat{H}(M + L)}.$$

Proof By substituting the lower-bound in Eq. (8.33) into Eq. (8.26), we obtain

$$\Theta \geq -\sigma^2 + \frac{\widehat{H}(L + M)\sigma^2 + \sum_{h=\widehat{H}+1}^L \gamma_h^2}{LM}.$$

This implies that $\Theta > 0$ unless the following hold:

$$\begin{aligned}\widehat{H} &< \frac{LM}{L + M} = \frac{L}{1 + \alpha}, \\ \sigma^2 &\geq \frac{\sum_{h=\widehat{H}+1}^L \gamma_h^2}{LM - \widehat{H}(L + M)}.\end{aligned}$$

Therefore, no local minimum exists if either of these conditions is violated. This completes the proof of Lemma 8.10. \square

It holds that

$$\frac{\sum_{h=\widehat{H}+1}^L \gamma_h^2}{LM - \widehat{H}(M + L)} \geq \frac{\sum_{h=\widehat{H}+1}^L \gamma_h^2}{M(L - \widehat{H})}, \quad (8.34)$$

of which the right-hand side is decreasing with respect to \widehat{H} . Combining Lemmas 8.5, 8.6, 8.7, and 8.10 and Eq. (8.34) completes the proof of Theorem 8.2. Corollary 8.3 is easily obtained from Lemmas 8.6 and 8.8.

8.4 Performance Analysis

To analyze the behavior of the EVB solution in the fully observed MF model, we rely on the *random matrix theory* (Marčenko and Pastur, 1967; Wachter, 1978; Johnstone, 2001; Bouchaud and Potters, 2003; Hoyle and Rattray, 2004; Baik and Silverstein, 2006), which describes the distribution of the singular values of random matrices in the limit when the matrix size goes to infinity. We first introduce some results obtained in the random matrix theory and then apply them to our analysis.

8.4.1 Random Matrix Theory

Assume that the observed matrix V is generated from the *spiked covariance model* (Johnstone, 2001):

$$V = U^* + \mathcal{E}, \quad (8.35)$$

where $\mathbf{U}^* \in \mathbb{R}^{L \times M}$ is a *true* signal matrix with rank H^* and singular values $\{\gamma_h^*\}_{h=1}^{H^*}$, and $\mathcal{E} \in \mathbb{R}^{L \times M}$ is a random matrix such that each element is independently drawn from a distribution with mean zero and variance σ^{*2} (not necessarily Gaussian). As the observed singular values $\{\gamma_h\}_{h=1}^L$ of \mathbf{V} , the true singular values $\{\gamma_h^*\}_{h=1}^{H^*}$ are also assumed to be arranged in the nonincreasing order.

We define normalized versions of the observed and the true singular values:

$$y_h = \frac{\gamma_h^2}{M\sigma^{*2}} \quad \text{for } h = 1, \dots, L, \quad (8.36)$$

$$\nu_h^* = \frac{\gamma_h^{*2}}{M\sigma^{*2}} \quad \text{for } h = 1, \dots, H^*. \quad (8.37)$$

In other words, $\{y_h\}_{h=1}^L$ are the eigenvalues of $\mathbf{V}\mathbf{V}^\top/(M\sigma^{*2})$, and $\{\nu_h^*\}_{h=1}^{H^*}$ are the eigenvalues of $\mathbf{U}^*\mathbf{U}^{*\top}/(M\sigma^{*2})$. Note the difference between x_h , defined by Eq. (6.132), and y_h : x_h is the squared observed singular value normalized with the model noise variance σ^2 , which is to be estimated, while y_h is the one normalized with the true noise variance σ^{*2} .

Define the empirical distribution of the observed eigenvalues $\{y_h\}_{h=1}^L$ by

$$p(y) = \frac{1}{L} \sum_{h=1}^L \delta(y - y_h), \quad (8.38)$$

where $\delta(y)$ denotes the Dirac delta function. When $H^* = 0$, the observed matrix $\mathbf{V} = \mathcal{E}$ consists only of noise, and its singular value distribution in the *large-scale limit* is specified by the following proposition:

Proposition 8.11 (*Marčenko and Pastur, 1967; Wachter, 1978*) *In the large-scale limit when L and M go to infinity with its ratio $\alpha = L/M$ fixed, the empirical distribution of the eigenvalue y of $\mathcal{E}\mathcal{E}^\top/(M\sigma^{*2})$ almost surely converges to*

$$p(y) \rightarrow p^{\text{MP}}(y) \equiv \frac{\sqrt{(y - \underline{y})(\bar{y} - y)}}{2\pi\alpha y} \theta(\underline{y} < y < \bar{y}), \quad (8.39)$$

$$\text{where } \bar{y} = (1 + \sqrt{\alpha})^2, \quad \underline{y} = (1 - \sqrt{\alpha})^2, \quad (8.40)$$

and $\theta(\cdot)$ is the indicator function, defined in Theorem 8.1.³

Figure 8.3 shows Eq. (8.39), which we call the *Marčenko–Pastur (MP) distribution*, for $\alpha = 0.1, 1$. The mean $\langle y \rangle_{p^{\text{MP}}(y)} = 1$ (which is constant for

³ Convergence is in weak topology in distribution, i.e., $p(y)$ almost surely converges to $p^{\text{MP}}(y)$ so that $\int f(y)p(y)dy = \int f(y)p^{\text{MP}}(y)dy$ for any bounded continuous function $f(y)$.

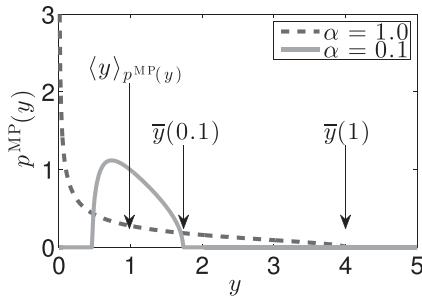


Figure 8.3 Marčenko–Pastur distribution.

any $0 < \alpha \leq 1$) and the upper-limits $\bar{y} = \bar{y}(\alpha)$ of the support for $\alpha = 0.1, 1$ are indicated by arrows. Proposition 8.11 states that the probability mass is concentrated in the range between $\underline{y} \leq y \leq \bar{y}$. Note that the MP distribution appears for a *single* sample matrix; differently from standard “large-sample” theories, Proposition 8.11 does not require one to average over many sample matrices.⁴ This single-sample property of the MP distribution is highly useful in our analysis because the MF model usually assumes a single observed matrix \mathbf{V} . We call the (unnormalized) singular value corresponding to the upper-limit \bar{y} , i.e.,

$$\bar{\gamma}^{\text{MPUL}} = \sqrt{M\sigma^{*2} \cdot \bar{y}} = (\sqrt{L} + \sqrt{M})\sigma^*, \quad (8.41)$$

the *Marčenko–Pastur upper limit (MPUL)*.

When $H^* > 0$, the true signal matrix \mathbf{U}^* affects the singular value distribution of \mathbf{V} . However, if $H^* \ll L$, the distribution can be approximated by a mixture of spikes (delta functions) and the MP distribution $p^{\text{MP}}(y)$. Let $H^{**} (\leq H^*)$ be the number of singular values of \mathbf{U}^* greater than $\gamma_h^* > \alpha^{1/4} \sqrt{M}\sigma^*$, i.e.,

$$\nu_{H^{**}}^* > \sqrt{\alpha} \quad \text{and} \quad \nu_{H^{**}+1}^* \leq \sqrt{\alpha}. \quad (8.42)$$

Then, the following proposition holds:

Proposition 8.12 (Baik and Silverstein, 2006) *In the large-scale limit when L and M go to infinity with finite α and H^* , it almost surely holds that*

$$y_h = y_h^{\text{Sig}} \equiv \left(1 + \nu_h^*\right) \left(1 + \frac{\alpha}{\nu_h^*}\right) \quad \text{for} \quad h = 1, \dots, H^{**}, \quad (8.43)$$

$$y_{H^{**}+1} = \bar{y}, \quad \text{and} \quad y_L = \underline{y}.$$

⁴ This property is called *self-averaging* (Bouchaud and Potters, 2003).

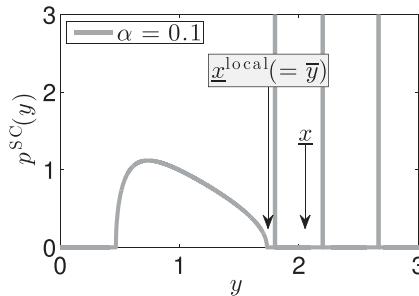


Figure 8.4 Spiked covariance distribution when $\{\nu_h^*\}_{h=1}^{H^{**}} = \{1.5, 1.0, 0.5\}$.

Furthermore, Hoyle and Rattray (2004) argued that, when L and M are large (but finite) and $H^* \ll L$, the empirical distribution of the eigenvalue y of $\mathbf{V}\mathbf{V}^\top/(M\sigma^2)$, is accurately approximated by

$$p(y) \approx p^{\text{SC}}(y) \equiv \frac{1}{L} \sum_{h=1}^{H^{**}} \delta(y - y_h^{\text{Sig}}) + \frac{L - H^{**}}{L} p^{\text{MP}}(y). \quad (8.44)$$

Figure 8.4 shows Eq. (8.44), which we call the *spiked covariance (SC) distribution*, for $\alpha = 0.1$, $H^{**} = 3$, and $\{\nu_h^*\}_{h=1}^{H^{**}} = \{1.5, 1.0, 0.5\}$. The SC distribution is irrespective of $\{\nu_h^*\}_{h=H^{**}+1}^{H^*}$, which satisfy $0 < \nu_h^* \leq \sqrt{\alpha}$ (see the definition (8.42) of H^{**}).

Proposition 8.12 states that in the large-scale limit, the large signal components such that $\nu_h^* > \sqrt{\alpha}$ appear outside the support of the MP distribution as spikes, while the other small signals are indistinguishable from the MP distribution (note that Eq. (8.43) implies that $y_h^{\text{Sig}} > \bar{y}$ for $\nu_h^* > \sqrt{\alpha}$). This implies that any PCA method fails to recover the true dimensionality, unless

$$\nu_{H^*}^* > \sqrt{\alpha}. \quad (8.45)$$

The condition (8.45) requires that \mathbf{U}^* has no small positive singular value such that $0 < \nu_h^* \leq \sqrt{\alpha}$, and therefore $H^{**} = H^*$.

The approximation (8.44) allows us to investigate more practical situations where the matrix size is finite. In Sections 8.4.2 and 8.4.4, respectively, we provide two theorems: one is based on Proposition 8.12 and guarantees perfect rank (PCA dimensionality) recovery of EVB learning in the large-scale limit, and the other one assumes that the approximation (8.44) exactly holds and provides a more realistic condition for perfect recovery.

8.4.2 Perfect Rank Recovery Condition in Large-Scale Limit

Now, we are almost ready for clarifying the behavior of the EVB solution. We assume that the model rank is set to be large enough, i.e., $H^* \leq H \leq L$, and all model parameters including the noise variance are estimated from observation (i.e., complete EVB learning). The last proposition on which our analysis relies is related to the property, called the *strong unimodality*, of the *log-concave distributions*:

Proposition 8.13 (*Ibragimov, 1956; Dharmadhikari and Joag-Dev, 1988*)
The convolution

$$g(s) = \langle f(s+t) \rangle_{p(t)} = \int f(s+t)p(t)dt$$

is quasiconvex, if $p(t)$ is a log-concave distribution, and $f(t)$ is a quasiconvex function.

In the large-scale limit, the summation over $h = 1, \dots, L$ in the objective $\mathcal{Q}(\sigma^{-2})$, given by Eq. (8.4), for noise variance estimation can be replaced with the expectation over the MP distribution $p^{\text{MP}}(y)$. By scaling variables, the objective can be written as a convolution with a scaled version of the MP distribution, which turns out to be log-concave. Accordingly, we can use Proposition 8.13 to show that $\mathcal{Q}(\sigma^{-2})$ is quasiconvex, which means that the noise variance estimation by EVB learning can be accurately performed by a local search algorithm. Combining this result with Proposition 8.12, we obtain the following theorem:

Theorem 8.14 *In the large-scale limit when L and M go to infinity with finite α and H^* , EVB learning almost surely recovers the true rank, i.e., $\widehat{H}^{\text{EVB}} = H^*$, if and only if*

$$\nu_{H^*}^* \geq \underline{\tau}, \quad (8.46)$$

where $\underline{\tau}$ is defined in Theorem 6.13.

Furthermore, the following corollary completely describes the behavior of the EVB solution in the large-scale limit:

Corollary 8.15 *In the large-scale limit, the objective $\mathcal{Q}(\sigma^{-2})$, defined by Eq. (8.4), for the noise variance estimation converges to a quasiconvex function, and it almost surely holds that*

$$\widehat{\tau}_h^{\text{EVB}} \left(\equiv \frac{\gamma_h \widehat{\gamma}_h^{\text{EVB}}}{M \widehat{\sigma}^2 \text{EVB}} \right) = \begin{cases} \nu_h^* & \text{if } \nu_h^* \geq \underline{\tau}, \\ 0 & \text{otherwise,} \end{cases} \quad (8.47)$$

$$\widehat{\sigma}^2 \text{EVB} = \sigma^{*2}.$$

One may get intuition of Eqs. (8.46) and (8.47) by comparing Eqs. (8.8) and (6.134) with Eq. (8.43): The estimator τ_h has the same relation to the observation x_h as the true signal v_h^* , and hence is an unbiased estimator of the signal. However, Theorem 8.14 does not even approximately hold in practical situations with moderate-sized matrices (see the numerical validation in Section 8.5). After proving Theorem 8.14 and Corollary 8.15, we will derive a more practical condition for perfect recovery in Section 8.4.4.

8.4.3 Proofs of Theorem 8.14 and Corollary 8.15

In the large-scale limit, we can substitute the expectation $\langle f(y) \rangle_{p(y)}$ for the summation $L^{-1} \sum_{h=1}^L f(y_h)$. We can also substitute the MP distribution $p^{\text{MP}}(y)$ for $p(y)$ for the expectation, since the contribution from the H^* signal components converges to zero. Accordingly, our objective (8.4) converges to

$$\begin{aligned} \Omega(\sigma^{-2}) &\rightarrow \Omega^{\text{LSL}}(\sigma^{-2}) \equiv \int_{\kappa}^{\bar{y}} \psi(\sigma^{*2} \sigma^{-2} y) p^{\text{MP}}(y) dy + \int_{\underline{y}}^{\kappa} \psi_0(\sigma^{*2} \sigma^{-2} y) p^{\text{MP}}(y) dy \\ &= \Omega^{\text{LSL-Full}}(\sigma^{-2}) - \int_{\max(x\sigma^2/\sigma^{*2}, \underline{y})}^{\kappa} \psi_1(\sigma^{*2} \sigma^{-2} y) p^{\text{MP}}(y) dy, \end{aligned} \quad (8.48)$$

$$\text{where } \Omega^{\text{LSL-Full}}(\sigma^{-2}) \equiv \int_{\underline{y}}^{\bar{y}} \psi(\sigma^{*2} \sigma^{-2} y) p^{\text{MP}}(y) dy, \quad (8.49)$$

and κ is a constant satisfying

$$\frac{H}{L} = \int_{\kappa}^{\bar{y}} p^{\text{MP}}(y) dy \quad (\underline{y} \leq \kappa \leq \bar{y}). \quad (8.50)$$

Note that \underline{x} , \underline{y} , and \bar{y} are defined by Eqs. (8.8) and (8.40), and it holds that

$$\underline{x} > \bar{y}. \quad (8.51)$$

We first investigate Eq. (8.49), which corresponds to the objective for the full-rank model (i.e., $H = L$). Let

$$\begin{aligned} s &= \log(\sigma^{-2}), \\ t &= \log y \quad \left(dt = \frac{1}{y} dy \right). \end{aligned}$$

Then Eq. (8.49) is written as a convolution:

$$\begin{aligned} \widetilde{\Omega}^{\text{LSL-Full}}(s) &\equiv \Omega^{\text{LSL-Full}}(e^s) = \int \psi(\sigma^{*2} e^{s+t}) e^t p^{\text{MP}}(e^t) dt \\ &= \int \widetilde{\psi}(s+t) p^{\text{LSMP}}(t) dt, \end{aligned} \quad (8.52)$$

where

$$\begin{aligned}\widetilde{\psi}(s) &= \psi(\sigma^{*2}e^s), \\ p^{\text{LSMP}}(t) &= e^t p^{\text{MP}}(e^t) \\ &= \frac{\sqrt{(e^t - \underline{y})(\bar{y} - e^t)}}{2\pi\alpha} \theta(\underline{y} < e^t < \bar{y}).\end{aligned}\quad (8.53)$$

Since Lemma 8.4 states that $\psi(x)$ is quasiconvex, its composition $\widetilde{\psi}(s)$ with the nondecreasing function $\sigma^{*2}e^s$ is also quasiconvex.

The following holds for $p^{\text{LSMP}}(t)$, which we call a log-scaled MP (LSMP) distribution:

Lemma 8.16 *The LSMP distribution (8.53) is log-concave.*

Proof Focusing on the support,

$$\log \underline{y} < t < \log \bar{y},$$

of the LSMP distribution (8.53), we define

$$\begin{aligned}f(t) &\equiv 2 \log p^{\text{LSMP}}(t) = 2 \log \frac{\sqrt{(e^t - \underline{y})(\bar{y} - e^t)}}{2\pi\alpha} \\ &= \log(-e^{2t} + (\underline{y} + \bar{y})e^t - \underline{y}\bar{y}) + \text{const.}\end{aligned}$$

Let

$$u(t) \equiv (e^t - \underline{y})(\bar{y} - e^t) = -e^{2t} + (\underline{y} + \bar{y})e^t - \underline{y}\bar{y} > 0, \quad (8.54)$$

and let

$$\begin{aligned}v(t) &\equiv \frac{\partial u}{\partial t} = -2e^{2t} + (\underline{y} + \bar{y})e^t = u - e^{2t} + \underline{y}\bar{y}, \\ w(t) &\equiv \frac{\partial^2 u}{\partial t^2} = -4e^{2t} + (\underline{y} + \bar{y})e^t = v - 2e^{2t},\end{aligned}$$

be the first and the second derivatives of u . Then, the first and the second derivatives of $f(t)$ are given by

$$\begin{aligned}\frac{\partial f}{\partial t} &= \frac{v}{u}, \\ \frac{\partial^2 f}{\partial t^2} &= \frac{uw - v^2}{u^2} \\ &= -\frac{e^t \left((\underline{y} + \bar{y})e^{2t} - 4\underline{y}\bar{y}e^t + (\underline{y} + \bar{y})\underline{y}\bar{y} \right)}{u^2}\end{aligned}$$

$$= -\frac{e^t(\underline{y} + \bar{y})}{u^2} \left(\left(e^t - \frac{2\underline{y}\bar{y}}{(\underline{y} + \bar{y})} \right)^2 + \frac{\underline{y}\bar{y}(\bar{y} - \underline{y})^2}{(\underline{y} + \bar{y})^2} \right) \\ \leq 0.$$

This proves the log-concavity of the LSMP distribution $p^{\text{LSMP}}(t)$, and completes the proof of Lemma 8.16. \square

Lemma 8.16 and Proposition 8.13 imply that $\tilde{\Omega}^{\text{LSL-Full}}(s)$ is quasiconvex, and therefore its composition $\Omega^{\text{LSL-Full}}(\sigma^{-2})$ with the nondecreasing function $\log(\sigma^{-2})$ is quasiconvex. The minimizer of $\Omega^{\text{LSL-Full}}(\sigma^{-2})$ can be found by evaluating the derivative Θ , given by Eq. (8.26), in the large-scale limit:

$$\begin{aligned} \Theta^{\text{Full}} \rightarrow \Theta^{\text{LSL-Full}} &= -\sigma^2 + \sigma^{*2} \int_{\underline{y}}^{\bar{y}} y \cdot p^{\text{MP}}(y) dy \\ &\quad - \int_{\underline{x}\sigma^2/\sigma^{*2}}^{\bar{y}} \tau(\sigma^{*2}\sigma^{-2}y; \alpha) p^{\text{MP}}(y) dy. \end{aligned} \quad (8.55)$$

Here, we used Eqs. (6.133) and (8.9). In the range

$$0 < \sigma^{-2} < \frac{x\sigma^{*2}}{\bar{y}} \quad \left(\text{i.e., } \frac{x\sigma^2}{\sigma^{*2}} > \bar{y} \right), \quad (8.56)$$

the third term in Eq. (8.55) is zero. Therefore, Eq. (8.55) is increasing with respect to σ^{-2} , and zero when

$$\sigma^2 = \sigma^{*2} \int_{\underline{y}}^{\bar{y}} y \cdot p^{\text{MP}}(y) dy = \sigma^{*2}.$$

Accordingly, $\Omega^{\text{LSL-Full}}(\sigma^{-2})$ is strictly convex in the range (8.56). Eq. (8.51) implies that the point $\sigma^{-2} = \sigma^{*2}$ is contained in the region (8.56), and therefore it is a local minimum of $\Omega^{\text{LSL-Full}}(\sigma^{-2})$. Combined with the quasi-convexity of $\Omega^{\text{LSL-Full}}(\sigma^{-2})$, we have the following lemma:

Lemma 8.17 *The objective $\Omega^{\text{LSL-Full}}(\sigma^{-2})$ for the full-rank model $H = L$ in the large-scale limit is quasiconvex with its minimizer at $\sigma^{-2} = \sigma^{*2}$. It is strictly convex in the range (8.56).*

For any κ ($\underline{y} < \kappa < \bar{y}$), the second term in Eq. (8.48) is zero in the range (8.56), which includes its minimizer at $\sigma^{-2} = \sigma^{*2}$. Since Lemma 8.4 states that $\psi_1(x)$ is decreasing for $x > \underline{x}$, the second term in Eq. (8.48) is nondecreasing in the region where

$$\left(\sigma^{*2} < \right) \frac{x\sigma^{*2}}{\bar{y}} \leq \sigma^{-2} < \infty.$$

Therefore, the quasi-convexity of $\mathcal{Q}^{\text{LSL-Full}}$ is inherited to \mathcal{Q}^{LSL} :

Lemma 8.18 *The objective $\mathcal{Q}^{\text{LSL}}(\sigma^{-2})$ for noise variance estimation in the large-scale limit is quasiconvex with its minimizer at $\sigma^{-2} = \sigma^{*-2}$. $\mathcal{Q}^{\text{LSL}}(\sigma^{-2})$ is strictly convex in the range (8.56).*

Thus we have proved that EVB learning accurately estimates the noise variance in the large-scale limit:

$$\widehat{\sigma}^2 \text{EVB} = \sigma^{*2}. \quad (8.57)$$

Assume that Eq. (8.45) holds. Then Proposition 8.12 guarantees that, in the large-scale limit, the following hold:

$$\frac{\gamma_{H^*}^2}{M\sigma^{*2}} \equiv y_{H^*} = (1 + \nu_{H^*}^*) \left(1 + \frac{\alpha}{\nu_{H^*}^*} \right), \quad (8.58)$$

$$\frac{\gamma_{H^*+1}^2}{M\sigma^{*2}} \equiv y_{H^*+1} = \bar{y} = (1 + \sqrt{\alpha})^2. \quad (8.59)$$

Remember that the EVB threshold is given by Eq. (8.8), i.e.,

$$\frac{(\gamma^{\text{EVB}})^2}{M\widehat{\sigma}^2 \text{EVB}} \equiv \underline{x} = \left(1 + \underline{\tau} \right) \left(1 + \frac{\alpha}{\underline{\tau}} \right). \quad (8.60)$$

Since Lemma 8.18 states that $\widehat{\sigma}^2 \text{EVB} = \sigma^{*2}$, comparing Eqs. (8.58) and (8.59) with Eq. (8.60) results in the following lemma:

Lemma 8.19 *It almost surely holds that*

$$\begin{aligned} \gamma_{H^*} &\geq \underline{\gamma}^{\text{EVB}} & \text{if and only if} && \nu_{H^*}^* \geq \underline{\tau}, \\ \gamma_{H^*+1} &< \underline{\gamma}^{\text{EVB}} & \text{for any} && \{\nu_h^*\}. \end{aligned} \quad (8.61)$$

This completes the proof of Theorem 8.14. Comparing Eqs. (6.134) and (8.43) under Lemmas 8.18 and 8.19 proves Corollary 8.15. \square

8.4.4 Practical Condition for Perfect Rank Recovery

Theorem 8.14 rigorously holds in the large-scale limit. However, it does not describe the behavior of the EVB solution very accurately in practical finite matrix-size cases. We can obtain a more practical condition for perfect recovery by relying on the approximation (8.44). We can prove the following theorem:

Theorem 8.20 *Let*

$$\xi = \frac{H^*}{L}$$

be the relevant rank ratio, and assume that

$$p(y) = p^{\text{SC}}(y). \quad (8.62)$$

Then, EVB learning recovers the true rank, i.e., $\widehat{H}^{\text{EVB}} = H^*$, if the following two inequalities hold:

$$\xi < \frac{1}{x}, \quad (8.63)$$

$$v_{H^*}^* > \frac{\left(\frac{x-1}{1-\underline{x}\xi} - \alpha\right) + \sqrt{\left(\frac{x-1}{1-\underline{x}\xi} - \alpha\right)^2 - 4\alpha}}{2}, \quad (8.64)$$

where \underline{x} is defined by Eq. (8.8).

Note that, in the large-scale limit, ξ converges to zero, and the sufficient condition, Eqs. (8.63) and (8.64), in Theorem 8.20 is equivalent to the necessary and sufficient condition (8.46) in Theorem 8.14.

Theorem 8.20 only requires that the SC distribution (8.44) well approximates the observed singular value distribution. Accordingly, it well describes the dependency of the EVB solution on ξ , which will be shown in numerical validation in Section 8.5. Theorem 8.20 states that, if the true rank H^* is small enough compared with L and the smallest signal $v_{H^*}^*$ is large enough, EVB learning perfectly recovers the true rank.

The following corollary also supports EVB learning:

Corollary 8.21 *Under the assumption (8.62) and the conditions (8.63) and (8.64), the objective $\Omega(\sigma^{-2})$ for the noise variance estimation has no local minimum (no stationary point if $\xi > 0$) that results in a wrong estimated rank $\widehat{H}^{\text{EVB}} \neq H^*$.*

This corollary states that, although the objective function (8.4) is nonconvex and possibly multimodal in general, any local minimum leads to the correct estimated rank. Therefore, perfect recovery does not require global search, but only local search, for noise variance estimation, if L and M are sufficiently large so that we can warrant Eq. (8.62).

In the next section, we give the proofs of Theorem 8.20 and Corollary 8.21, and then show numerical experiments that support the theory.

8.4.5 Proofs of Theorem 8.20 and Corollary 8.21

We regroup the terms in Eq. (8.4) as follows:

$$\Omega(\sigma^{-2}) = \Omega_1(\sigma^{-2}) + \Omega_0(\sigma^{-2}), \quad (8.65)$$

where

$$\mathcal{Q}_1(\sigma^{-2}) = \frac{1}{H^*} \sum_{h=1}^{H^*} \psi\left(\frac{\gamma_h^2}{M} \sigma^{-2}\right), \quad (8.66)$$

$$\mathcal{Q}_0(\sigma^{-2}) = \frac{1}{L - H^*} \left(\sum_{h=H^*+1}^H \psi\left(\frac{\gamma_h^2}{M} \sigma^{-2}\right) + \sum_{h=H+1}^L \psi_0\left(\frac{\gamma_h^2}{M} \sigma^{-2}\right) \right). \quad (8.67)$$

In the following, assuming that Eq. (8.62) holds and

$$y_{H^*} > \bar{y}, \quad (8.68)$$

we derive a sufficient condition for any local minimizer to lie only in $\sigma^{-2} \in \mathcal{B}_{H^*}$, with which Lemma 8.7 proves Theorem 8.20.

Under the assumption (8.62) and the condition (8.68), $\mathcal{Q}_0(\sigma^{-2})$, defined by Eq. (8.67), is equivalent to the objective $\mathcal{Q}^{\text{LSL}}(\sigma^{-2})$ in the large-scale limit. Using Lemma 8.18, and noting that

$$\underline{\sigma}_{H^*+1}^{-2} = \frac{M_x}{\gamma_{H^*+1}}^2 = \frac{x\sigma^{*-2}}{\bar{y}} > \sigma^{*-2}, \quad (8.69)$$

we have the following lemma:

Lemma 8.22 $\mathcal{Q}_0(\sigma^{-2})$ is quasiconvex with its minimizer at

$$\sigma^{-2} = \sigma^{*-2}.$$

$\mathcal{Q}_0(\sigma^{-2})$ is strictly convex in the range

$$0 < \sigma^{-2} < \underline{\sigma}_{H^*+1}^{-2}.$$

Using Lemma 8.22 and the strict quasiconvexity of $\psi(x)$, we can deduce the following lemma:

Lemma 8.23 $\mathcal{Q}(\sigma^{-2})$ is nondecreasing (increasing if $\xi > 0$) in the range $\underline{\sigma}_{H^*+1}^2 < \sigma^{-2} < \infty$.

Proof Lemma 8.22 states that $\mathcal{Q}_0(\sigma^{-2})$, defined by Eq. (8.67), is quasiconvex with its minimizer at

$$\sigma^{-2} = \left(\frac{\sum_{h=H^*+1}^L \gamma_h^2}{(L - H^*)M} \right)^{-1} = \sigma^{*-2}.$$

Since $\mathcal{Q}_1(\sigma^{-2})$, defined by Eq. (8.66), is the sum of strictly quasiconvex functions with their minimizers at $\sigma^{-2} = M/\gamma_h^2 < \sigma^{*-2}$ for $h = 1, \dots, H^*$, our objective $\mathcal{Q}(\sigma^{-2})$, given by Eq. (8.65), is nondecreasing (increasing if $H^* > 0$) for

$$\sigma^{-2} \geq \sigma^{*-2}.$$

Since Eq. (8.69) implies that $\underline{\sigma}_{H^*+1}^{-2} > \sigma^{*-2}$, $\Omega(\sigma^{-2})$ is nondecreasing (increasing if $\xi > 0$) for $\sigma^{-2} > \underline{\sigma}_{H^*+1}^{-2}$, which completes the proof of Lemma 8.23. \square

Using the bounds given by Eq. (8.33) and Lemma 8.22, we also obtain the following lemma:

Lemma 8.24 $\Omega(\sigma^{-2})$ is increasing at $\sigma^{-2} = \underline{\sigma}_{H^*+1}^2 - 0$.⁵ It is decreasing at $\sigma^{-2} = \underline{\sigma}_{H^*}^2 + 0$ if the following hold:

$$\xi < \frac{1}{(1 + \sqrt{\alpha})^2}, \quad (8.70)$$

$$y_{H^*} > \frac{x(1 - \xi)}{1 - \xi(1 + \sqrt{\alpha})^2}. \quad (8.71)$$

Proof Lemma 8.22 states that $\Omega_0(\sigma^{-2})$ is strictly convex in the range $0 < \sigma^{-2} < \underline{\sigma}_{H^*+1}^2$, and minimized at $\sigma^{-2} = \sigma^{*-2}$. Since Eq. (8.69) implies that $\sigma^{*-2} < \underline{\sigma}_{H^*+1}^2$, $\Omega_0(\sigma^{-2})$ is increasing at $\sigma^{-2} = \underline{\sigma}_{H^*+1}^2 - 0$. Since $\Omega_1(\sigma^{-2})$ is the sum of strictly quasiconvex functions with their minimizers at $\sigma^{-2} = M/\gamma_h^2 < \sigma^{*-2}$ for $h = 1, \dots, H^*$, $\Omega(\sigma^{-2})$ is also increasing at $\sigma^{-2} = \underline{\sigma}_{H^*+1}^2 - 0$.

Let us investigate the sign of the derivative Θ of $\Omega(\sigma^{-2})$ at $\sigma^{-2} = \underline{\sigma}_{H^*}^2 + 0 \in \mathcal{B}_{H^*}$. Substituting the upper-bound in Eq. (8.33) into Eq. (8.26), we have

$$\begin{aligned} \Theta &< -\sigma^2 + \frac{H^*(\sqrt{L} + \sqrt{M})^2\sigma^2 + \sum_{h=H^*+1}^L \gamma_h^2}{LM} \\ &= -\sigma^2 + \frac{H^*(\sqrt{L} + \sqrt{M})^2\sigma^2 + (L - H^*)M\sigma^{*2}}{LM}. \end{aligned} \quad (8.72)$$

The right-hand side of Eq. (8.72) is negative if the following hold:

$$\xi = \frac{H^*}{L} < \frac{M}{(\sqrt{L} + \sqrt{M})^2} = \frac{1}{(1 + \sqrt{\alpha})^2}, \quad (8.73)$$

$$\sigma^2 > \frac{(L - H^*)M\sigma^{*2}}{LM - H^*(\sqrt{L} + \sqrt{M})^2} = \frac{(1 - \xi)\sigma^{*2}}{1 - \xi(1 + \sqrt{\alpha})^2}. \quad (8.74)$$

Assume that the first condition (8.73) holds. Then the second condition (8.74) holds at $\sigma^{-2} = \underline{\sigma}_{H^*}^2 + 0$, if

$$\underline{\sigma}_{H^*}^{-2} < \frac{1 - \xi(1 + \sqrt{\alpha})^2}{(1 - \xi)}\sigma^{*-2},$$

⁵ By “ -0 ” we denote an arbitrarily large negative value.

or equivalently,

$$y_{H^*} = \frac{\gamma_{H^*}^2}{M\sigma^{*2}} = \underline{x} \cdot \frac{\sigma_{H^*}^2}{\sigma^{*2}} > \frac{\underline{x}(1-\xi)}{1-\xi(1+\sqrt{\alpha})^2},$$

which completes the proof of Lemma 8.24. \square

Finally, we obtain the following lemma:

Lemma 8.25 $\Omega(\sigma^{-2})$ is decreasing in the range $0 < \sigma^{-2} < \underline{\sigma}_{H^*}^2$ if the following hold:

$$\xi < \frac{1}{\underline{x}}, \quad (8.75)$$

$$y_{H^*} > \frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}. \quad (8.76)$$

Proof In the range $0 < \sigma^{-2} < \underline{\sigma}_{H^*}^2$, the estimated rank (8.27) is bounded as

$$0 \leq \widehat{H} \leq H^* - 1. \quad (8.77)$$

Substituting the upper-bound in Eq. (8.33) into Eq. (8.26), we have

$$\begin{aligned} \Theta &< -\sigma^2 + \frac{\widehat{H}(\sqrt{L} + \sqrt{M})^2\sigma^2 + \sum_{h=\widehat{H}+1}^{H^*} \gamma_h^2 + \sum_{h=H^*+1}^L \gamma_h^2}{LM} \\ &= -\sigma^2 + \frac{\widehat{H}(\sqrt{L} + \sqrt{M})^2\sigma^2 + \sum_{h=\widehat{H}+1}^{H^*} \gamma_h^2 + (L-H^*)M\sigma^{*2}}{LM}. \end{aligned} \quad (8.78)$$

The right-hand side of Eq. (8.78) is negative, if the following hold:

$$\frac{\widehat{H}}{L} < \frac{M}{(\sqrt{L} + \sqrt{M})^2} = \frac{1}{(1 + \sqrt{\alpha})^2}, \quad (8.79)$$

$$\sigma^2 > \frac{\sum_{h=\widehat{H}+1}^{H^*} \gamma_h^2 + (L-H^*)M\sigma^{*2}}{LM - \widehat{H}(\sqrt{L} + \sqrt{M})^2}. \quad (8.80)$$

Assume that

$$\xi = \frac{H^*}{L} < \frac{1}{(1 + \sqrt{\alpha})^2}.$$

Then both of the conditions (8.79) and (8.80) hold for $\sigma^{-2} \in (0, \underline{\sigma}_{H^*}^2)$, if the following holds:

$$\underline{\sigma}_{H+1}^{-2} < \frac{LM - \widehat{H}(\sqrt{L} + \sqrt{M})^2}{\sum_{h=\widehat{H}+1}^{H^*} \gamma_h^2 + (L-H^*)M\sigma^{*2}} \quad \text{for } \widehat{H} = 0, \dots, H^* - 1. \quad (8.81)$$

Since the sum $\sum_{h=\widehat{H}+1}^{H^*} \gamma_h^2$ in the right-hand side of Eq. (8.81) is upper-bounded as

$$\sum_{h=\widehat{H}+1}^{H^*} \gamma_h^2 \leq (H^* - \widehat{H})\gamma_{\widehat{H}+1}^2,$$

Eq. (8.81) holds if

$$\begin{aligned} \underline{\sigma}_{\widehat{H}+1}^{-2} &< \frac{LM - \widehat{H}(\sqrt{L} + \sqrt{M})^2}{(H^* - \widehat{H})\gamma_{\widehat{H}+1}^2 + (L - H^*)L\sigma^{*2}} \\ &= \frac{1 - \frac{\widehat{H}}{L}(1 + \sqrt{\alpha})^2}{(\xi - \frac{\widehat{H}}{L})\frac{\gamma_{\widehat{H}+1}^2}{M} + (1 - \xi)\sigma^{*2}} \quad \text{for } \widehat{H} = 0, \dots, H^* - 1. \end{aligned} \quad (8.82)$$

Using Eq. (8.24), the condition (8.82) is rewritten as

$$\begin{aligned} \frac{\gamma_{\widehat{H}+1}^2}{M\underline{x}} &> \frac{(\xi - \frac{\widehat{H}}{L})\frac{\gamma_{\widehat{H}+1}^2}{M} + (1 - \xi)\sigma^{*2}}{1 - \frac{\widehat{H}}{L}(1 + \sqrt{\alpha})^2} \\ \left(1 - \frac{\widehat{H}}{L}(1 + \sqrt{\alpha})^2\right) \frac{\gamma_{\widehat{H}+1}^2}{M\sigma^{*2}} &> \left(\xi\underline{x} - \frac{\widehat{H}}{L}\underline{x}\right) \frac{\gamma_{\widehat{H}+1}^2}{M\sigma^{*2}} + (1 - \xi)\underline{x}, \end{aligned}$$

or equivalently

$$y_{\widehat{H}+1} = \frac{\gamma_{\widehat{H}+1}^2}{M\sigma^{*2}} > \frac{(1 - \xi)\underline{x}}{\left(1 - \xi\underline{x} + \frac{\widehat{H}}{L}(\underline{x} - (1 + \sqrt{\alpha})^2)\right)} \quad \text{for } \widehat{H} = 0, \dots, H^* - 1. \quad (8.83)$$

Note that $\underline{x} > \bar{y} = (1 + \sqrt{\alpha})^2$. Further bounding both sides, we have the following sufficient condition for Eq. (8.83) to hold:

$$y_{H^*} > \frac{(1 - \xi)\underline{x}}{\max(0, 1 - \xi\underline{x})}. \quad (8.84)$$

Thus we obtain the conditions (8.75) and (8.76) for Θ to be negative for $\sigma^{-2} \in (0, \underline{\sigma}_{H^*}^2)$, which completes the proof of Lemma 8.25. \square

Lemmas 8.23, 8.24, and 8.25 together state that, if all the conditions (8.68) and (8.70) through (8.76) hold, at least one local minimum exists in the correct range $\sigma^{-2} \in \mathcal{B}_{H^*}$, and no local minimum (no stationary point if $\xi > 0$) exists outside the correct range. Therefore, we can estimate the correct rank $\widehat{H}^{\text{EVB}} = H^*$ by using a local search algorithm for noise variance estimation. Choosing the tightest conditions, we have the following lemma:

Lemma 8.26 $\mathcal{Q}(\sigma^{-2})$ has a global minimum in $\sigma^{-2} \in \mathcal{B}_{H^*}$, and no local minimum (no stationary point if $\xi > 0$) outside \mathcal{B}_{H^*} , if the following hold:

$$\begin{aligned}\xi &< \frac{1}{x}, \\ y_{H^*} &= \frac{\gamma_{H^*}^2}{M\sigma^{*2}} > \frac{x(1-\xi)}{1-x\xi}.\end{aligned}\quad (8.85)$$

Using Eq. (8.43), Eq. (8.85) can be written with the *true* signal amplitude as follows:

$$(1 + \nu_{H^*}^*) \left(1 + \frac{\alpha}{\nu_{H^*}^*} \right) - \frac{x(1-\xi)}{1-x\xi} > 0. \quad (8.86)$$

The left-hand side of Eq. (8.86) can be factorized as follows:

$$\begin{aligned}\frac{1}{\nu_{H^*}^*} \left(\nu_{H^*}^* - \frac{\left(\frac{x(1-\xi)}{1-x\xi} - (1+\alpha) \right) + \sqrt{\left(\frac{x(1-\xi)}{1-x\xi} - (1+\alpha) \right)^2 - 4\alpha}}{2} \right) \\ \cdot \left(\nu_{H^*}^* - \frac{\left(\frac{x(1-\xi)}{1-x\xi} - (1+\alpha) \right) - \sqrt{\left(\frac{x(1-\xi)}{1-x\xi} - (1+\alpha) \right)^2 - 4\alpha}}{2} \right) > 0.\end{aligned}\quad (8.87)$$

When Eq. (8.45) holds, the last factor in the left-hand side in Eq. (8.87) is positive. Therefore, we have the following condition:

$$\begin{aligned}\nu_{H^*}^* &> \frac{\left(\frac{x(1-\xi)}{1-x\xi} - (1+\alpha) \right) + \sqrt{\left(\frac{x(1-\xi)}{1-x\xi} - (1+\alpha) \right)^2 - 4\alpha}}{2} \\ &= \frac{\left(\frac{x-1}{1-x\xi} - \alpha \right) + \sqrt{\left(\frac{x-1}{1-x\xi} - \alpha \right)^2 - 4\alpha}}{2}.\end{aligned}\quad (8.88)$$

Lemma 8.26 with the condition (8.85) replaced with the condition (8.88) leads to Theorem 8.20 and Corollary 8.21.

8.5 Numerical Verification

Figure 8.5 shows numerical simulation results for $M = 200$ and $L = 20, 100, 200$. \mathcal{E} was drawn from the independent Gaussian distribution with mean 0 and variance $\sigma^{*2} = 1$, and *true* signal singular values $\{\gamma_h^*\}_{h=1}^{H^*}$ were

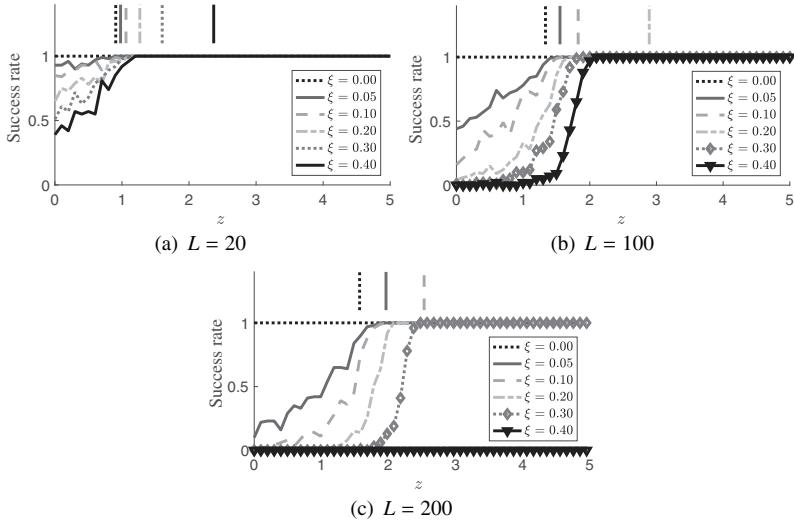


Figure 8.5 Success rate of rank recovery in numerical simulation for $M = 200$. The horizontal axis indicates the lower limit of the support of the simulated true signal distribution, i.e., $z \approx \sqrt{V_H}$. The recovery condition (8.64) for finite-sized matrices is indicated by a vertical bar with the same line style for each ξ . The leftmost vertical bar, which corresponds to the condition (8.64) for $\xi = 0$, coincides with the recovery condition (8.46) for infinite-sized matrices.

drawn from the uniform distribution on $[z\sqrt{M}\sigma^*, 10\sqrt{M}\sigma^*]$ for different z , which is indicated by the horizontal axis. We used Algorithm 16, which will be introduced in Chapter 9, to compute the global EVB solution.

The vertical axis indicates the success rate of rank recovery over 100 trials, i.e., the proportion of the trials giving $\widehat{H}^{\text{EVB}} = H^*$. If the condition (8.63) on ξ is violated, the corresponding curve is depicted with markers. Otherwise, the condition (8.64) on $v_{H^*}^* (= \gamma_{H^*}^{*2}/(M\sigma^{*2}))$ is indicated by a vertical bar with the same line style for each ξ . In other words, Theorem 8.20 states that the success rate should be equal to one if $z (> \gamma_{H^*}^*/(\sqrt{M}\sigma^{*2}))$ is larger than the value indicated by the vertical bar. The leftmost vertical bar, which corresponds to the condition (8.64) for $\xi = 0$, coincides with the recovery condition (8.46), given by Theorem 8.14, for infinite-sized matrices.

We see that Theorem 8.20 with the condition (8.64) approximately holds for these moderate-sized matrices, while Theorem 8.14 with the condition (8.46), which does not depend on the relevant rank ratio ξ , immediately breaks for positive ξ .

8.6 Comparison with Laplace Approximation

Here, we compare EVB learning with an alternative dimensionality selection method (Hoyle, 2008) based on the *Laplace approximation (LA)*. Consider the PCA application, where D denotes the dimensionality of the observation space, and N denotes the number of samples, i.e., in our MF notation to keep $L \leq M$,

$$\begin{aligned} L = D, M = N &\quad \text{if} \quad D \leq N, \\ L = N, M = D &\quad \text{if} \quad D > N. \end{aligned} \quad (8.89)$$

Right after Tipping and Bishop (1999) proposed the *probabilistic PCA*, Bishop (1999a) proposed to select the PCA dimensionality by maximizing the marginal likelihood:

$$p(\mathbf{V}) = \langle p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \rangle_{p(\mathbf{A})p(\mathbf{B})}. \quad (8.90)$$

Since the marginal likelihood (8.90) is computationally intractable, he approximated it by LA, and suggested Gibbs sampling and VB learning as alternatives. The VB variant, of which the model is almost the same as the MF defined by Eqs. (6.1) through (6.3), was also proposed by himself (Bishop, 1999b) along with a standard local solver similar to Algorithm 1 in Chapter 3.

The LA-based approach was polished in Minka (2001a), by introducing a conjugate prior⁶ on \mathbf{B} to $p(\mathbf{V}|\mathbf{B}) = \langle p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \rangle_{p(\mathbf{A})}$, and ignoring the non-leading terms that do not grow fast as the number N of samples goes to infinity. Hoyle (2008) pointed out that Minka's method is inaccurate when $D \gg N$, and proposed the *overlap (OL) method*, a further polished variant of the LA-based approach. A notable difference of the OL method from most of the LA-based methods is that the OL method applies LA around a more accurate estimator than the MAP estimator.⁷ Thanks to the use of the accurate estimator, the OL method behaves *optimally* in the large-scale limit when D and N go to infinity, while Minka's method does not. We will clarify the meaning of the optimality and discuss it in more detail in Section 8.7.

The OL method minimizes an approximation to the negative logarithm of the marginal likelihood (8.90), which depends on estimators for $\lambda_h = b_h^2 + \sigma^2$ and σ^2 computed by an iterative algorithm, over the hypothetical model rank $H = 0, \dots, L$ (see Appendix C for the detailed computational procedure). Figure 8.6 shows numerical simulation results that compare EVB learning and the OL method: Figure 8.6(a) shows the success rate for the no-signal case $\xi = 0$ ($H^* = 0$), while Figures 8.6(b) through 8.6(f) show the success rate for $\xi = 0.05$ and $D = 20, 100, 200, 400$, and $1,000$, respectively. We also show

⁶ This conjugate prior does not satisfy the implicit requirement, footnoted in Section 1.2.4, that the moments of the family member can be computed analytically.

⁷ As explained in Section 2.2.1, LA is usually applied around the MAP estimator.

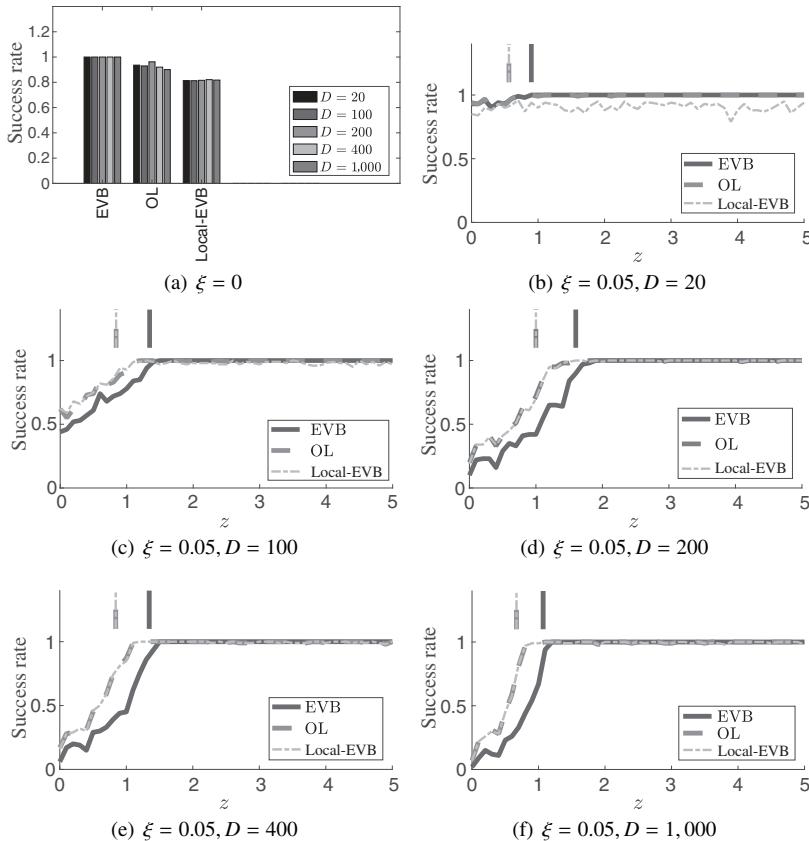


Figure 8.6 Success rate of PCA dimensionality recovery by (global) EVB learning, the OL method, and the local-EVB estimator for $N = 200$. Vertical bars indicate the recovery conditions, Eq. (8.46) for EVB learning, and Eq. (8.95) for the OL method and the local-EVB estimator, in the large-scale limit.

the performance of the *local-EVB estimator* (6.131), which was computed by a local solver (Algorithm 18 introduced in Chapter 9). For the OL method and the local-EVB estimator, we initialized the noise variance estimator to $10^{-4} \cdot \sum_{h=1}^L \gamma_h^2 / (LM)$.

In comparison with the OL method, EVB learning shows its conservative nature: It exhibits almost zero false positive rate (Figure 8.6(a)) at the expense of low sensitivity (Figures 8.6(c) through 8.6(f)). Actually, because of its low sensitivity, EVB learning does not behave optimally in the large-scale limit. The local-EVB estimator, on the other hand, behaves similarly to the OL method, for which the reason will be elucidated in the next section.

8.7 Optimality in Large-Scale Limit

Consider the large-scale limit, and assume that the model rank H is set to be large enough but finite so that $H \geq H^*$ and $H/L \rightarrow 0$. Then the rank estimation procedure, detailed in Appendix C, by the OL method is reduced to counting the number of components such that $\widehat{\lambda}_h^{\text{OL-LSL}} > \widehat{\sigma}^2 \text{OL-LSL}$, i.e.,

$$\widehat{H}^{\text{OL-LSL}} = \sum_{h=1}^L \theta\left(\widehat{\lambda}_h^{\text{OL-LSL}} > \widehat{\sigma}^2 \text{OL-LSL}\right), \quad (8.91)$$

where $\theta(\cdot)$ is the indicator function defined in Theorem 8.1. Here $\widehat{\lambda}_h^{\text{OL-LSL}}$ and $\widehat{\sigma}^2 \text{OL-LSL}$ are computed by iterating the following updates until convergence:

$$\widehat{\lambda}_h^{\text{OL-LSL}} = \begin{cases} \check{\lambda}_h^{\text{OL-LSL}} & \text{if } \gamma_h \geq \underline{\gamma}^{\text{local-EVB}}, \\ \widehat{\sigma}^2 \text{OL-LSL} & \text{otherwise,} \end{cases} \quad (8.92)$$

$$\widehat{\sigma}^2 \text{OL-LSL} = \frac{1}{(M-H)} \left(\sum_{l=1}^L \frac{\gamma_l^2}{L} - \sum_{h=1}^H \widehat{\lambda}_h^{\text{OL-LSL}} \right), \quad (8.93)$$

$$\text{where } \check{\lambda}_h^{\text{OL-LSL}} = \frac{\gamma_h^2}{2L} \left(1 - \frac{(M-L)\widehat{\sigma}^2 \text{OL-LSL}}{\gamma_h^2} \right. \\ \left. + \sqrt{\left(1 - \frac{(M-L)\widehat{\sigma}^2 \text{OL-LSL}}{\gamma_h^2} \right)^2 - \frac{4L\widehat{\sigma}^2 \text{OL-LSL}}{\gamma_h^2}} \right). \quad (8.94)$$

The OL method evaluates its objective, which approximates the negative logarithm of the marginal likelihood (8.90), after the updates (8.92) and (8.93) converge for each hypothetical H , and adopts the minimizer $\widehat{H}^{\text{OL-LSL}}$ as the rank estimator. However, Hoyle (2008) proved that, in the large-scale limit, the objective decreases as H increases, as long as Eq. (8.94) is a real number (or equivalently $\gamma_h \geq \underline{\gamma}^{\text{local-EVB}}$ holds) for all $h = 1, \dots, H$ at the convergence. Accordingly, Eq. (8.91) holds.

Interestingly, the threshold in Eq. (8.92) coincides with the local-EVB threshold (6.127). Moreover, the updates (8.92) and (8.93) for the OL method are equivalent to the updates (9.29) and (9.30) for the local-EVB estimator (Algorithm 18) with the following correspondence:

$$\widehat{\lambda}_h^{\text{OL-LSL}} = \frac{\gamma_h \widehat{\gamma}_h^{\text{local-EVB}}}{L} + \widehat{\sigma}^2 \text{local-EVB}, \\ \widehat{\sigma}^2 \text{OL-LSL} = \widehat{\sigma}^2 \text{local-EVB}.$$

Thus, the rank estimation procedure by the OL method and that by the local-EVB estimator are equivalent, and therefore $\widehat{H}^{\text{OL-LSL}} = \widehat{H}^{\text{local-EVB}}$ in the large-scale limit.

If the noise variance is accurately estimated, i.e., $\widehat{\sigma}^2 = \sigma^{*2}$, the threshold $\underline{\gamma}_{\text{local-EVB}}^{\text{OL}}$ both for the OL method and the local-EVB estimator coincides with the MPUL (8.41), which corresponds to the minimum detectable observed singular value. By using this fact, the optimality of the OL method in the large-scale limit was shown:

Proposition 8.27 (Hoyle, 2008) *In the large-scale limit, when L and M go to infinity with finite α , H^* , and $H (\geq H^*)^8$, the OL method almost surely recovers the true rank, i.e., $\widehat{H}_{\text{OL-LSL}}^{\text{OL-LSL}} = H^*$, if and only if*

$$\nu_{H^*}^* > \sqrt{\alpha}. \quad (8.95)$$

It almost surely holds that

$$\begin{aligned} \frac{\widehat{\lambda}_h^{\text{OL-LSL}}}{\widehat{\sigma}^2 \text{OL-LSL}} - 1 &= \nu_h^*, \\ \widehat{\sigma}^2 \text{OL-LSL} &= \sigma^{*2}. \end{aligned}$$

The condition (8.95) coincides with the condition (8.45), which any PCA method requires for perfect dimensionality recovery. In this sense, the OL method, as well as the local-EVB estimator, is optimal in the large-scale limit.

On the other hand, Theorem 8.14 implies that (global) EVB learning is not optimal in the large-scale limit but more conservative (see the difference between $\underline{\tau}$ and $\sqrt{\alpha}$ in Figure 6.4). In Figure 8.6, the conditions for perfect dimensionality recovery in the large-scale limit are indicated by vertical bars:

$$z = \sqrt{\underline{\tau}} \text{ for EVB, and } z = \sqrt{\tau^{\text{local}}} = \alpha^{1/4} \text{ for OL and local-EVB.}$$

All methods accurately estimate the noise variance in the large-scale limit, i.e.,

$$\widehat{\sigma}^2 \text{EVB} = \widehat{\sigma}^2 \text{OL-LSL} = \widehat{\sigma}^2 \text{local-EVB} = \sigma^{*2}.$$

Taking this into account, we indicate the recovery conditions in Figure 8.4 by arrows at

$$y = \underline{x} \text{ for EVB, and } y = \underline{x}^{\text{local}} (= \bar{y}) \text{ for OL and local-EVB,}$$

respectively. Figure 8.4 implies that, in this particular case, EVB learning discards the third spike coming from the third true signal $\nu_3^* = 0.5$, while the OL method and the local-EVB estimator successfully capture it as a signal.

When the matrix size is finite, the conservative nature of EVB learning is not always bad, since it offers almost zero false positive rate, which makes

⁸ Unlike our analysis in Section 8.4, Hoyle (2008) assumed $H/L \rightarrow 0$ to prove that the OL method accurately estimates the noise variance.

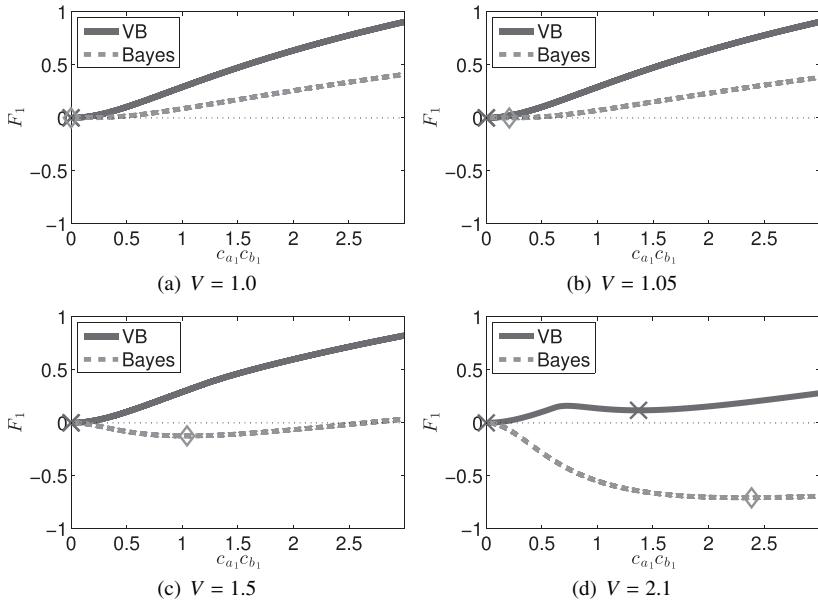


Figure 8.7 The VB free energy contribution (6.55) from the (first) component and its counterpart (8.96) of Bayesian learning for $L = M = H = 1$ and $\sigma^2 = 1$. Markers indicate the local minima.

Theorem 8.20 approximately hold for finite cases, as seen in Figures 8.5 and 8.6. However, the fact that not (global) EVB learning but the local-EVB estimator is optimal in the large-scale limit might come from inaccurate approximation to the Bayes posterior by the VB posterior. Having this in mind, we discuss the difference between VB learning and full Bayesian learning in the remainder of this section.

Figure 8.7 shows the VB free energy contribution (6.55) from the (first) component as a function of $c_a c_b$, and its counterpart of Bayesian learning:

$$2F_1^{\text{Bayes}} = -2 \log \langle p(V|A, B) \rangle_{p(A)p(B)} - \left(\log(2\pi\sigma^2) + \frac{V^2}{\sigma^2} \right), \quad (8.96)$$

which was numerically computed. We see that the minimizer (shown as a diamond) of the Bayes free energy is at $c_a c_b \rightarrow +0$ until V exceeds 1.

The difference in behavior between EVB learning and the local-EVB estimator appears in the nonempty range of the observed value V where the positive local solution exists but gives positive free energy. Figure 8.7(d) shows this case, where a bump exists between two local minima (indicated by crosses). On the other hand, such multimodality is not observed in empirical

full Bayesian learning (see the dashed curves in Figures 8.7(a) through 8.7(d)). We can say that this multimodality in EVB learning with a bump between two local minima is induced by the independence constraint for VB learning. We further guess that it is this bump that pushes the EVB threshold from the optimal point (at the local-EVB threshold) to a larger value. Further investigation is necessary to fully understand this phenomenon.

9

Global Solver for Matrix Factorization

The analytic-form solutions, derived in Chapter 6, for VB learning and EVB learning in fully observed MF can naturally be used to develop efficient and reliable VB solvers. Some properties, shown in Chapter 8, can also be incorporated when the noise variance is unknown and to be estimated.

In this chapter, we introduce global solvers for VB learning and EVB learning (Nakajima et al., 2013a, 2015), and how to extend them to more general cases with missing entries and nonconjugate likelihoods (Seeger and Bouchard, 2012).

9.1 Global VB Solver for Fully Observed MF

We consider the MF model, introduced in Section 3.1:

$$p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{V} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2\right), \quad (9.1)$$

$$p(\mathbf{A}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{A}\mathbf{C}_A^{-1}\mathbf{A}^\top)\right), \quad (9.2)$$

$$p(\mathbf{B}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{B}\mathbf{C}_B^{-1}\mathbf{B}^\top)\right), \quad (9.3)$$

where $\mathbf{V} \in \mathbb{R}^{L \times M}$ is an observed matrix;

$$\begin{aligned} \mathbf{A} &= (\mathbf{a}_1, \dots, \mathbf{a}_H) = (\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_M)^\top \in \mathbb{R}^{M \times H}, \\ \mathbf{B} &= (\mathbf{b}_1, \dots, \mathbf{b}_H) = (\tilde{\mathbf{b}}_1, \dots, \tilde{\mathbf{b}}_L)^\top \in \mathbb{R}^{L \times H}, \end{aligned}$$

are parameter matrices; and \mathbf{C}_A , \mathbf{C}_B , and σ^2 are hyperparameters. The prior covariance hyperparameters are restricted to be diagonal:

$$\begin{aligned}\mathbf{C}_A &= \text{Diag}(c_{a_1}^2, \dots, c_{a_H}^2), \\ \mathbf{C}_B &= \text{Diag}(c_{b_1}^2, \dots, c_{b_H}^2).\end{aligned}$$

Our VB solver gives the global solution to the following minimization problem,

$$\widehat{r} = \underset{r}{\operatorname{argmin}} F(r) \quad \text{s.t.} \quad r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}), \quad (9.4)$$

of the free energy

$$F(r) = \left\langle \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{p(\mathbf{V}|\mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}. \quad (9.5)$$

Assume that $L \leq M$ without loss of generality, and let

$$\mathbf{V} = \sum_{h=1}^L \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top \quad (9.6)$$

be the singular value decomposition (SVD) of the observed matrix $\mathbf{V} \in \mathbb{R}^{L \times M}$. According to Theorem 6.7, the VB solution is given by

$$\widehat{\mathbf{U}}^{\text{VB}} = \widehat{\mathbf{B}}\widehat{\mathbf{A}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \quad \text{where} \quad \widehat{\gamma}_h^{\text{VB}} = \begin{cases} \check{\gamma}_h^{\text{VB}} & \text{if } \gamma_h \geq \underline{\gamma}_h^{\text{VB}}, \\ 0 & \text{otherwise,} \end{cases} \quad (9.7)$$

for

$$\underline{\gamma}_h^{\text{VB}} = \sigma \sqrt{\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2} + \sqrt{\left(\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2}\right)^2 - LM}}, \quad (9.8)$$

$$\check{\gamma}_h^{\text{VB}} = \gamma_h \left(1 - \frac{\sigma^2}{2\underline{\gamma}_h^{\text{VB}}} \left(M + L + \sqrt{(M-L)^2 + \frac{4\underline{\gamma}_h^{\text{VB}}}{c_{a_h}^2 c_{b_h}^2}} \right) \right). \quad (9.9)$$

Corollary 6.8 completely specifies the VB posterior, which is written as

$$r(\mathbf{A}, \mathbf{B}) = \prod_{h=1}^H \text{Gauss}_M(\mathbf{a}_h; \widehat{a}_h \boldsymbol{\omega}_{a_h}, \widehat{\sigma}_{a_h}^2 \mathbf{I}_M) \prod_{h=1}^H \text{Gauss}_L(\mathbf{b}_h; \widehat{b}_h \boldsymbol{\omega}_{b_h}, \widehat{\sigma}_{b_h}^2 \mathbf{I}_L) \quad (9.10)$$

with the following variational parameters: if $\gamma_h > \underline{\gamma}_h^{\text{VB}}$,

$$\widehat{a}_h = \pm \sqrt{\check{\gamma}_h^{\text{VB}} \widehat{\delta}_h^{\text{VB}}}, \quad \widehat{b}_h = \pm \sqrt{\frac{\check{\gamma}_h^{\text{VB}}}{\widehat{\delta}_h^{\text{VB}}}}, \quad \widehat{\sigma}_{a_h}^2 = \frac{\sigma^2 \widehat{\delta}_h^{\text{VB}}}{\check{\gamma}_h^{\text{VB}}}, \quad \widehat{\sigma}_{b_h}^2 = \frac{\sigma^2}{\check{\gamma}_h^{\text{VB}} \widehat{\delta}_h^{\text{VB}}}, \quad (9.11)$$

Algorithm 15 Global VB solver for fully observed matrix factorization.

- 1: Transpose $\mathbf{V} \rightarrow \mathbf{V}^\top$ if $L > M$, and set H ($\leq L$) to a sufficiently large value.
 - 2: Compute the SVD (9.6) of \mathbf{V} .
 - 3: Apply Eqs. (9.7) through (9.9) to get the VB estimator.
 - 4: If necessary, compute the variational parameters by using Eqs. (9.11) through (9.14), which specify the VB posterior (9.10). We can also evaluate the free energy by using Eqs. (9.15) and (9.16).
-

where
$$\widehat{\delta}_h^{\text{VB}} \left(\equiv \frac{\widehat{a}_h}{\widehat{b}_h} \right) = \frac{c_{a_h}}{\sigma^2} \left(\gamma_h - \check{\gamma}_h^{\text{VB}} - \frac{L\sigma^2}{\gamma_h} \right), \quad (9.12)$$

and otherwise,

$$\widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = c_{a_h}^2 \left(1 - \frac{L\zeta_h^{\text{VB}}}{\sigma^2} \right), \quad \widehat{\sigma}_{b_h}^2 = c_{b_h}^2 \left(1 - \frac{M\zeta_h^{\text{VB}}}{\sigma^2} \right), \quad (9.13)$$

where

$$\zeta_h^{\text{VB}} \left(\equiv \widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2 \right) = \frac{\sigma^2}{2LM} \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} - \sqrt{\left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right)^2 - 4LM} \right). \quad (9.14)$$

The free energy can be written as

$$2F = LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^L \gamma_h^2}{\sigma^2} + \sum_{h=1}^H 2F_h, \quad (9.15)$$

where
$$2F_h = M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2}{c_{b_h}^2} - (L + M) + \frac{-2\widehat{a}_h \widehat{b}_h \gamma_h + (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2)(\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2)}{\sigma^2}. \quad (9.16)$$

Based on these results, we can straightforwardly construct a global solver for VB learning, which is given in Algorithm 15.

9.2 Global EVB Solver for Fully Observed MF

EVB learning, where the hyperparameters \mathbf{C}_A , \mathbf{C}_A , and σ^2 are also estimated from observation, solves the following minimization problem,

$$\widehat{r} = \underset{r, \mathbf{C}_A, \mathbf{C}_A, \sigma^2}{\operatorname{argmin}} F \quad \text{s.t.} \quad r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}), \quad (9.17)$$

of the free energy (9.5).

According to Theorem 6.13, given the noise variance σ^2 , the EVB solution can be written as

$$\widehat{\mathbf{U}}^{\text{EVB}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{EVB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \quad \text{where} \quad \widehat{\gamma}_h^{\text{EVB}} = \begin{cases} \check{\gamma}_h^{\text{EVB}} & \text{if } \gamma_h \geq \underline{\gamma}^{\text{EVB}}, \\ 0 & \text{otherwise,} \end{cases} \quad (9.18)$$

for

$$\begin{aligned} \underline{\gamma}^{\text{EVB}} &= \sigma \sqrt{M \left(1 + \underline{\tau}\right) \left(1 + \frac{\alpha}{\underline{\tau}}\right)}, \\ \check{\gamma}_h^{\text{EVB}} &= \frac{\gamma_h}{2} \left(1 - \frac{(M+L)\sigma^2}{\gamma_h^2} + \sqrt{\left(1 - \frac{(M+L)\sigma^2}{\gamma_h^2}\right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}}\right). \end{aligned}$$

Here

$$\alpha = \frac{L}{M} \quad (0 < \alpha \leq 1),$$

is the “squaredness” of the observed matrix \mathbf{V} , and $\underline{\tau} = \underline{\tau}(\alpha)$ is the unique zero-cross point of the following function:

$$\Xi(\tau; \alpha) = \Phi(\tau) + \Phi\left(\frac{\tau}{\alpha}\right), \quad \text{where} \quad \Phi(z) = \frac{\log(z+1)}{z} - \frac{1}{2}. \quad (9.19)$$

Summarizing Lemmas 6.14, 6.16, and 6.19, the EVB posterior is completely specified by Eq. (9.10) with the variational parameters given as follows: If $\gamma_h \geq \underline{\gamma}^{\text{EVB}}$,

$$\widehat{a}_h = \pm \sqrt{\check{\gamma}_h^{\text{EVB}} \widehat{\delta}_h^{\text{EVB}}}, \quad \widehat{b}_h = \pm \sqrt{\frac{\check{\gamma}_h^{\text{EVB}}}{\widehat{\delta}_h^{\text{EVB}}}}, \quad (9.20)$$

$$\widehat{\sigma}_{a_h}^2 = \frac{\sigma^2 \widehat{\delta}_h^{\text{EVB}}}{\gamma_h}, \quad \widehat{\sigma}_{b_h}^2 = \frac{\sigma^2}{\gamma_h \widehat{\delta}_h^{\text{EVB}}}, \quad c_{a_h} c_{b_h} = \sqrt{\frac{\gamma_h \check{\gamma}_h^{\text{EVB}}}{LM}}, \quad (9.21)$$

$$\text{where} \quad \widehat{\delta}_h^{\text{EVB}} = \sqrt{\frac{M \check{\gamma}_h^{\text{EVB}}}{L \gamma_h}} \left(1 + \frac{L \sigma^2}{\gamma_h \check{\gamma}_h^{\text{EVB}}}\right), \quad (9.22)$$

and otherwise

$$\widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = \sqrt{\zeta^{\text{EVB}}}, \quad \widehat{\sigma}_{b_h}^2 = \sqrt{\zeta^{\text{EVB}}}, \quad c_{a_h} c_{b_h} = \sqrt{\zeta^{\text{EVB}}}, \quad (9.23)$$

$$\text{where} \quad \widehat{\zeta}^{\text{EVB}} \rightarrow +0. \quad (9.24)$$

To use the preceding result, we need to prepare a table of $\underline{\tau}$ by computing the zero-cross point of Eq. (9.19) as a function of α . A simple approximation $\underline{\tau} \approx \underline{z} \sqrt{\alpha} \approx 2.5129 \sqrt{\alpha}$ is a reasonable alternative (see Figure 6.4).

For noise variance estimation, we can use Theorems 8.1 and 8.2, derived in Chapter 8. Specifically, after performing the SVD (9.6), we first estimate the noise variance by solving the following problem:

$$\widehat{\sigma}^2 \text{EVB} = \operatorname{argmin}_{\sigma^2} \Omega(\sigma^{-2}), \quad (9.25)$$

$$\text{s.t.} \quad \max \left(\frac{\underline{\sigma}_H^2}{\overline{H}+1}, \frac{\sum_{h=\overline{H}+1}^L \gamma_h^2}{M(L-\overline{H})} \right) \leq \sigma^2 \leq \frac{1}{LM} \sum_{h=1}^L \gamma_h^2, \quad (9.26)$$

where

$$\Omega(\sigma^{-2}) = \frac{1}{L} \left(\sum_{h=1}^H \psi \left(\frac{\gamma_h^2}{M\sigma^2} \right) + \sum_{h=\overline{H}+1}^L \psi_0 \left(\frac{\gamma_h^2}{M\sigma^2} \right) \right), \quad (9.27)$$

$$\psi(x) = \psi_0(x) + \theta(x > \underline{x}) \psi_1(x),$$

$$\psi_0(x) = x - \log x,$$

$$\psi_1(x) = \log(\tau(x; \alpha) + 1) + \alpha \log \left(\frac{\tau(x; \alpha)}{\alpha} + 1 \right) - \tau(x; \alpha),$$

$$\underline{x} = (1 + \underline{\tau}) \left(1 + \frac{\alpha}{\underline{\tau}} \right),$$

$$\tau(x; \alpha) = \frac{1}{2} \left(x - (1 + \alpha) + \sqrt{(x - (1 + \alpha))^2 - 4\alpha} \right),$$

$$\underline{\sigma}_h^2 = \begin{cases} \infty & \text{for } h = 0, \\ \frac{\gamma_h^2}{M\underline{x}} & \text{for } h = 1, \dots, L, \\ 0 & \text{for } h = L + 1, \end{cases}$$

$$\overline{H} = \min \left(\left\lceil \frac{L}{1 + \alpha} \right\rceil - 1, H \right).$$

Problem (9.25) is simply a one-dimensional search for the minimizer of the function $\Omega(\sigma^{-2})$, which is typically smooth. Note also that, if the matrix size is large enough, Corollary 8.21 states that any local minimizer is accurate enough to estimate the correct rank. Given the estimated noise variance $\sigma^2 = \widehat{\sigma}^2 \text{EVB}$, Eq. (9.18) gives the EVB solution.

Algorithm 16 summarizes the procedure explained in the preceding discussion. This algorithm gives the global solution, provided that the global solution to the one-dimensional search problem (9.25) is attained. If the noise variance σ^2 is known, we should simply skip Step 4.

Algorithm 16 Global EVB solver for fully observed matrix factorization.

- 1: Transpose $\mathbf{V} \rightarrow \mathbf{V}^\top$ if $L > M$, and set $H (\leq L)$ to a sufficiently large value.
 - 2: Refer to the table of $\underline{\tau}(\alpha)$ at $\alpha = L/M$ (or use a simple approximation $\underline{\tau} \approx 2.5129 \sqrt{\alpha}$).
 - 3: Compute the SVD (9.6) of \mathbf{V} .
 - 4: Solve the one-dimensional search problem (9.25) to get $\widehat{\sigma}^2 \text{EVB}$.
 - 5: Apply Eq. (9.18) to get the EVB estimator $\{\widehat{\gamma}_h^{\text{EVB}}\}_{h=1}^H$ for $\sigma^2 = \widehat{\sigma}^2 \text{EVB}$.
 - 6: If necessary, compute the variational parameters and the hyperparameters by using Eqs. (9.20) through (9.24), which specify the EVB posterior (9.10). We can also evaluate the free energy by using Eqs. (9.15) and (9.16), noting that $F_h \rightarrow +0$ for h such that $\gamma_h < \underline{\gamma}^{\text{EVB}}$.
-

Algorithm 17 Iterative EVB solver for fully observed matrix factorization.

- 1: Transpose $\mathbf{V} \rightarrow \mathbf{V}^\top$ if $L > M$, and set $H (\leq L)$ to a sufficiently large value.
 - 2: Refer to the table of $\underline{\tau}(\alpha)$ at $\alpha = L/M$ (or use a simple approximation $\underline{\tau} \approx 2.5129 \sqrt{\alpha}$).
 - 3: Compute the SVD (9.6) of \mathbf{V} .
 - 4: Initialize the noise variance $\widehat{\sigma}^2 \text{EVB}$ to the lower bound in Eq. (9.26).
 - 5: Apply Eq. (9.18) to update the EVB estimator $\{\widehat{\gamma}_h^{\text{EVB}}\}_{h=1}^H$.
 - 6: Apply Eq. (9.28) to update the noise variance estimator $\widehat{\sigma}^2 \text{EVB}$.
 - 7: Compute the variational parameters and the hyperparameters by using Eqs. (9.20) through (9.24).
 - 8: Evaluate the free energy (9.15), noting that $F_h \rightarrow +0$ for h such that $\gamma_h < \underline{\gamma}^{\text{EVB}}$.
 - 9: Iterate Steps 5 through 8 until convergence (until the energy decrease becomes smaller than a threshold).
-

Another implementation is to iterate Eq. (9.18) and

$$\widehat{\sigma}^2 \text{EVB} = \frac{1}{LM} \left(\sum_{l=1}^L \gamma_l^2 - \sum_{h=1}^H \gamma_h \widehat{\gamma}_h^{\text{EVB}} \right) \quad (9.28)$$

in turn. Note that Eq. (9.28) was derived in Corollary 8.3 and can be used as an update rule for the noise variance estimator, given the current EVB estimators $\{\widehat{\gamma}_h^{\text{EVB}}\}_{h=1}^H$. Although it is not guaranteed, this iterative algorithm (Algorithm 17) tends to converge to the global solution if we initialize the noise variance $\widehat{\sigma}^2 \text{EVB}$ to be sufficiently small (Nakajima et al., 2015). We recommend to initialize it to the lower-bound given in Eq. (9.26).

Algorithm 18 Local-EVB solver for fully observed matrix factorization.

-
- 1: Transpose $\mathbf{V} \rightarrow \mathbf{V}^\top$ if $L > M$, and set H ($\leq L$) to a sufficiently large value.
 - 2: Refer to the table of $\underline{\tau}(\alpha)$ at $\alpha = L/M$ (or use a simple approximation $\underline{\tau} \approx 2.5129 \sqrt{\alpha}$).
 - 3: Compute the SVD (9.6) of \mathbf{V} .
 - 4: Initialize the noise variance $\widehat{\sigma}^2 \text{ local-EVB}$ to the lower-bound in Eq. (9.26).
 - 5: Apply Eq. (9.29) to update the local-EVB estimator $\{\widehat{\gamma}_h^{\text{local-EVB}}\}_{h=1}^H$.
 - 6: Apply Eq. (9.30) to update the noise variance estimator $\widehat{\sigma}^2 \text{ local-EVB}$.
 - 7: Compute the variational parameters and the hyperparameters by using Eqs. (9.20) through (9.24).
 - 8: Evaluate the free energy (9.15), noting that $F_h \rightarrow +0$ for h such that $\gamma_h < \underline{\gamma}^{\text{local-EVB}}$.
 - 9: Iterate Steps 5 through 8 until convergence (until the energy decrease becomes smaller than a threshold).
-

Finally, we introduce an iterative solver, in Algorithm 18, for the *local-EVB estimator* (6.131), which iterates the following updates:

$$\widehat{\gamma}_h^{\text{local-EVB}} = \begin{cases} \widehat{\gamma}_h^{\text{EVB}} & \text{if } \gamma_h \geq \underline{\gamma}^{\text{local-EVB}}, \\ 0 & \text{otherwise,} \end{cases} \quad (9.29)$$

$$\widehat{\sigma}^2 \text{ local-EVB} = \frac{1}{LM} \left(\sum_{l=1}^L \gamma_l^2 - \sum_{h=1}^H \gamma_h \widehat{\gamma}_h^{\text{local-EVB}} \right), \quad (9.30)$$

where

$$\underline{\gamma}^{\text{local-EVB}} \equiv (\sqrt{L} + \sqrt{M}) \sigma \quad (9.31)$$

is the local-EVB threshold, defined by Eq. (6.127). If we initialize the noise variance $\widehat{\sigma}^2 \text{ local-EVB}$ to be sufficiently small, this algorithm tends to retain the positive local-EVB solution for each h if it exists, and therefore does not necessarily converge to the global EVB solution. The interesting relation between the local-EVB estimator and the overlap (OL) method (Hoyle, 2008), an alternative dimensionality selection method based on the Laplace approximation, was discussed in Section 8.7.

9.3 Empirical Comparison with the Standard VB Algorithm

Here we see how efficient the global solver (Algorithm 16) is in comparison with the standard VB algorithm (Algorithm 1 in Section 3.1) on artificial and benchmark data.

9.3.1 Experiment on Artificial Data

We first created an artificial data set (*Artificial1*) with the data matrix size $L = 100$ and $M = 300$, and the true rank $H^* = 20$. We randomly drew *true* matrices $\mathbf{A}^* \in \mathbb{R}^{M \times H^*}$ and $\mathbf{B}^* \in \mathbb{R}^{L \times H^*}$ so that each entry of \mathbf{A}^* and \mathbf{B}^* follows $\text{Gauss}_1(0, 1)$, where $\text{Gauss}_1(\mu, \sigma^2)$ denotes the one-dimensional Gaussian distribution with mean μ and variance σ^2 . An observed matrix \mathbf{V} was created by adding noise subject to $\text{Gauss}_1(0, 1)$ to each entry of $\mathbf{B}^* \mathbf{A}^{*\top}$.

We evaluated the performance under the complete empirical Bayesian scenario, where all variational parameters and hyperparameters are estimated from observation. We used the full-rank model (i.e., $H = \min(L, M)$), expecting that irrelevant $H - H^*$ components will be automatically trimmed out by the automatic relevance determination (ARD) effect (see Chapters 7 and 8).

We compare the global solver (Algorithm 16) and the standard VB algorithm (Algorithm 1 in Section 3.1), and show the free energy, the computation time, and the estimated rank over iterations in Figure 9.1. For the standard VB algorithm, initial values were set in the following way: $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ are randomly created so that each entry follows $\text{Gauss}_1(0, 1)$. Other variables are set to $\widehat{\Sigma}_A = \widehat{\Sigma}_B = \mathbf{C}_A = \mathbf{C}_B = \mathbf{I}_H$ and $\sigma^2 = 1$. Note that we rescale \mathbf{V} so that $\|\mathbf{V}\|_{\text{Fro}}^2 / (LM) = 1$, before starting iterations. We ran the standard algorithm 10

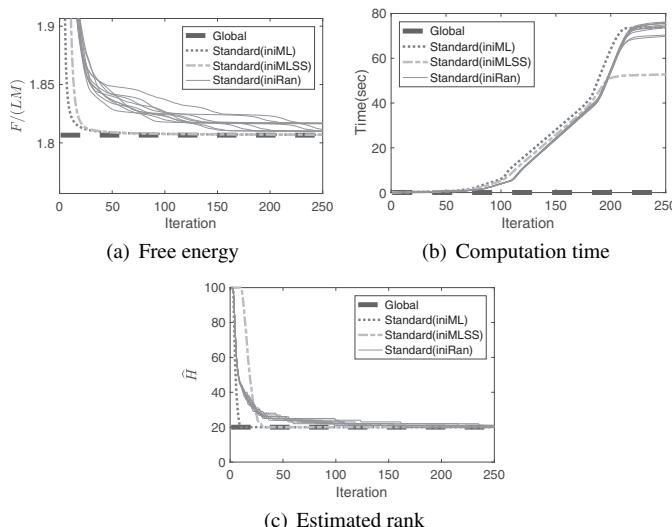


Figure 9.1 Experimental results on the *Artificial1* data, where the data matrix size is $L = 100$ and $M = 300$, and the true rank is $H^* = 20$.

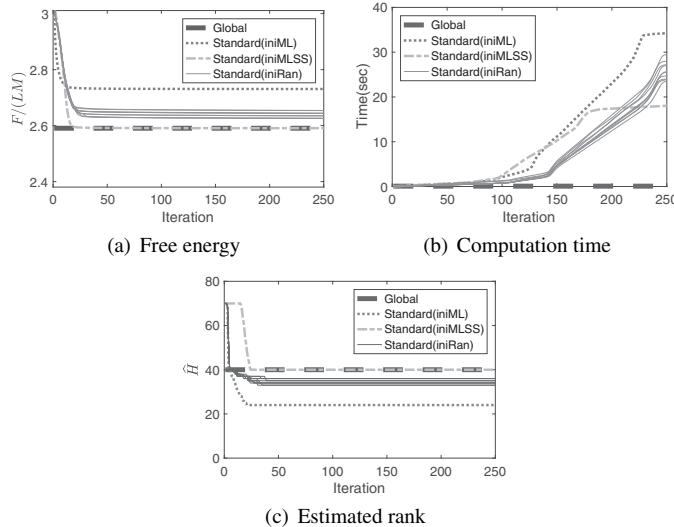


Figure 9.2 Experimental results on the *Artificial2* data set ($L = 70$, $M = 300$, and $H^* = 40$).

times, starting from different initial points, and each trial is plotted by a solid curve labeled as “Standard(iniRan)” in Figure 9.1.

The global solver has no iteration loop, and therefore the corresponding dashed line labeled as “Global” is constant over iterations. We see that the global solver finds the true rank $\hat{H} = H^* = 20$ immediately (~ 0.1 sec on average over 10 trials), while the standard iterative algorithm does not converge in 60 sec.

Figure 9.2 shows experimental results on another artificial data set (*Artificial2*) where $L = 70$, $M = 300$, and $H^* = 40$. In this case, all the 10 trials of the standard algorithm are trapped at local minima. We empirically observed that the local minimum problem tends to be more critical when H^* is large (close to H).

We also evaluated the standard algorithm with different initialization schemes. The curve labeled as “Standard(iniML)” indicates the standard algorithm starting from the maximum likelihood (ML) solution: $(\hat{\mathbf{a}}_h, \hat{\mathbf{b}}_h) = (\sqrt{\gamma_h} \omega_{a_h}, \sqrt{\gamma_h} \omega_{b_h})$. The initial values for other variables are the same as the random initialization. Figures 9.1 and 9.2 show that the ML initialization generally makes convergence faster than the random initialization, but suffers from the local minimum problem more severely—it tends to converge to a worse local minimum.

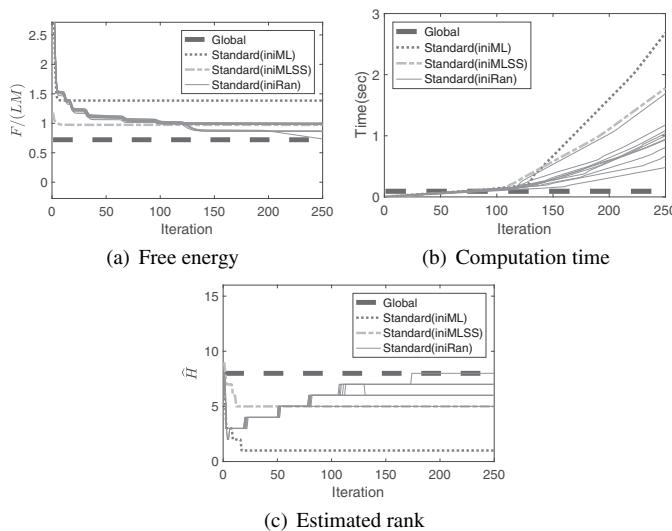


Figure 9.3 Experimental results on the *Glass* data set ($L = 9, M = 214$).

We observed that starting from a small noise variance tends to alleviate the local minimum problem at the expense of slightly slower convergence. The curve labeled as “Standard(iniMLSS)” indicates the standard algorithm starting from the ML solution with a small noise variance $\sigma^2 = 0.0001$. We see in Figures 9.1 and 9.2 that this initialization improves the quality of solutions, and successfully finds the true rank for these artificial data sets. However, we will show in Section 9.3.2 that this scheme still suffers from the local minimum problem on benchmark datasets.

9.3.2 Experiment on Benchmark Data

Figures 9.3 through 9.5 show the experimental results on the *Glass*, the *Satimage*, and the *Spectf* data sets available from the University of California, Irvine (UCI) repository (Asuncion and Newman, 2007). A similar tendency to the artificial data experiment (Figures 9.1 and 9.2) is observed: “Standard(iniRan)” converges slowly, and is often trapped at a local minimum with a *wrong* estimated rank;¹ “Standard(iniML)” converges slightly faster but to a worse local minimum; and “Standard(iniMLSS)” tends to give a better solution. Unlike the artificial data experiment, “Standard(iniMLSS)” fails to

¹ Since the *true* ranks of the benchmark data sets are unknown, we mean by a *wrong* rank a rank different from the one giving the lowest free energy.

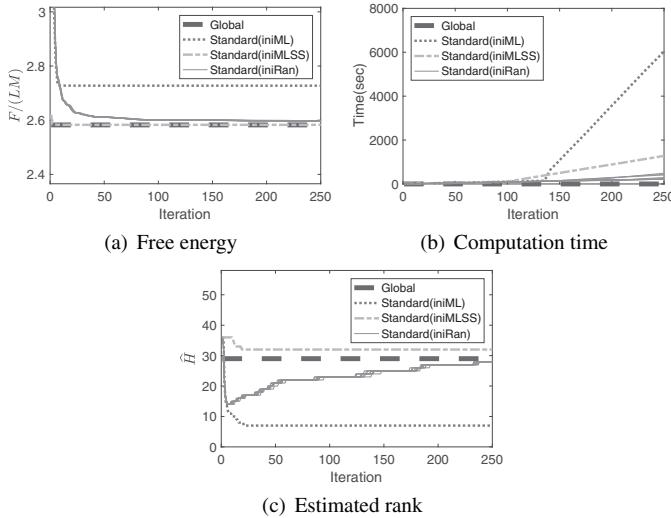


Figure 9.4 Experimental results on the *Satimage* data set ($L = 36, M = 6435$).

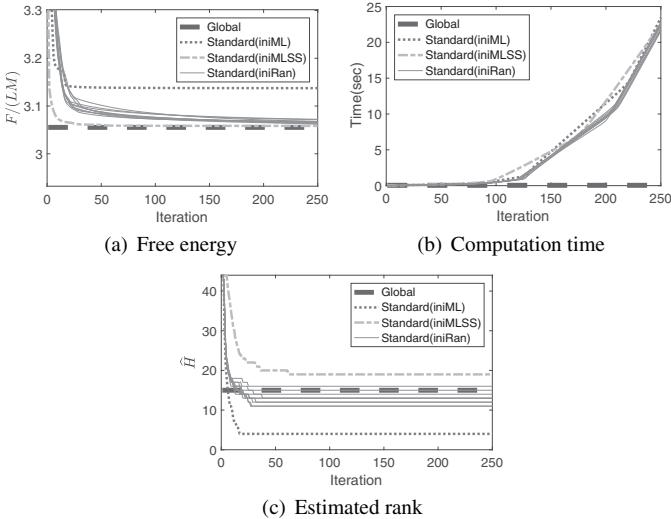


Figure 9.5 Experimental results on the *Spectf* data set ($L = 44, M = 267$).

find the *correct* rank in these benchmark data sets. We also conducted experiments on other benchmark data sets and found that the standard VB algorithm generally converges slowly, and sometimes suffers from the local minimum problem, while the global solver gives the global solution immediately.

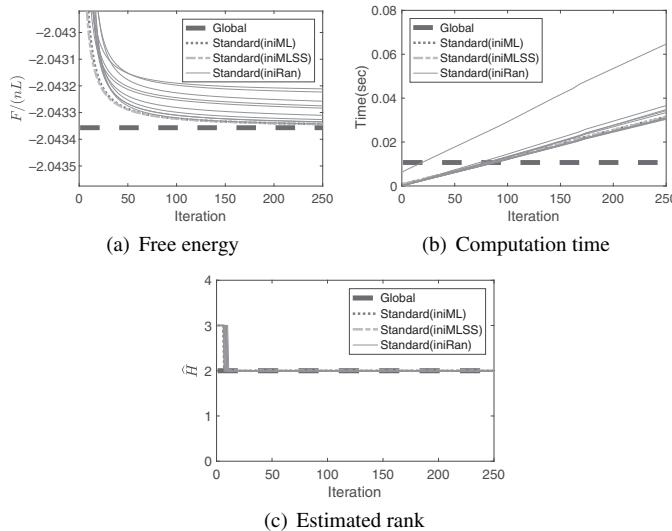


Figure 9.6 Experimental results on the *Concrete Slump Test* data set (an RRR task with $L = 3, M = 7$).

Finally, we applied EVB learning to the *reduced rank regression* (RRR) model (see Section 3.1.2), of which the model likelihood is given by Eq. (3.36). Figure 9.6 shows the results on the *Concrete Slump Test* data set, where we centered the $L = 3$ -dimensional outputs and prewhitened the $M = 7$ -dimensional inputs. We also standardized the outputs so that the variance of each element is equal to one. Note that we cannot directly apply Algorithm 16 for the RRR model. Instead, we use Algorithm 16 with a fixed noise variance (skipping Step 4) and apply one-dimensional search to minimize the free energy (3.42), in order to estimate the *rescaled* noise variance σ^2 . For the standard VB algorithm, the rescaled noise variance should be updated by Eq. (3.43), instead of Eq. (3.28). The original noise variance σ'^2 is recovered by Eq. (3.40) for both cases.

Overall, the global solver showed excellent performance over the standard VB algorithm.

9.4 Extension to Nonconjugate MF with Missing Entries

The global solvers introduced in Section 9.1 can be directly applied only for the fully observed isotropic Gaussian likelihood (9.1). However, the global solver can be used as a subroutine to develop efficient algorithms for more

general cases. In this section, we introduce the approach by Seeger and Bouchard (2012), where an *iterative singular value shrinkage* algorithm was proposed, based on the global VB solver (Algorithm 15) and local variational approximation (Section 2.1.7).

9.4.1 Nonconjugate MF Model

Consider the following model:

$$p(\mathbf{V}|\mathbf{A}, \mathbf{B}) = \prod_{l=1}^L \prod_{m=1}^M \phi_{l,m}(V_{l,m}, \tilde{\mathbf{b}}_l^\top \tilde{\mathbf{a}}_m), \quad (9.32)$$

$$p(\mathbf{A}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{A}\mathbf{C}_A^{-1}\mathbf{A}^\top)\right), \quad (9.33)$$

$$p(\mathbf{B}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{B}\mathbf{C}_B^{-1}\mathbf{B}^\top)\right), \quad (9.34)$$

where $\phi_{l,m}(v|u)$ is a function of v and u , and satisfy

$$-\frac{\partial^2 \log \phi_{l,m}(v, u)}{\partial u^2} \leq \frac{1}{\sigma^2} \quad (9.35)$$

for any l, m, v , and u .

The function $\phi_{l,m}(v, u)$ corresponds to the model distribution of the (l, m) th entry v of \mathbf{V} parameterized by u . If

$$\phi_{l,m}(v, u) = \text{Gauss}_1(v; u, \sigma^2) \quad (9.36)$$

for all l and m , the model (9.32) through (9.34) is reduced to the fully observed isotropic Gaussian MF model (9.1)–(9.3), and the Hessian,

$$-\frac{\partial^2 \log \phi_{l,m}(v, u)}{\partial u^2} = \frac{1}{\sigma^2},$$

of the negative log-likelihood is a constant with respect to v and u .

The model (9.32) through (9.34) can cover the case with missing entries by setting the noise variance in Eq. (9.36) to $\sigma^2 \rightarrow \infty$ for the unobserved entries (the condition (9.35) is tight for the smallest σ^2). Other one-dimensional distributions, including the *Bernoulli distribution* with *sigmoid* parameterization and the *Poisson distribution*, satisfy the condition (9.35) for a certain σ^2 , which will be introduced in Section 9.4.3.

9.4.2 Local Variational Approximation for Non-conjugate MF

The VB learning problem (9.4) minimizes the free energy, which can be written as

$$F(r) = \left\langle \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{\prod_{l=1}^L \prod_{m=1}^M \phi_{l,m}(V_{l,m}, \tilde{\mathbf{b}}_l^\top \tilde{\mathbf{a}}_m) p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}. \quad (9.37)$$

In order to make the global VB solver applicable as a subroutine, we instead solve the following joint minimization problem,

$$\hat{r} = \underset{r, \Xi}{\operatorname{argmin}} \bar{F}(r, \Xi) \quad \text{s.t.} \quad r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}), \quad (9.38)$$

of an upper-bound of the free energy,

$$F \leq \bar{F}(r, \Xi) \equiv \left\langle \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{\prod_{l=1}^L \prod_{m=1}^M \underline{\phi}_{l,m}(V_{l,m}, \tilde{\mathbf{b}}_l^\top \tilde{\mathbf{a}}_m, \Xi_{l,m}) p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}, \quad (9.39)$$

where

$$\underline{\phi}_{l,m}(v, u, \xi) \leq \phi_{l,m}(v, u) \quad (9.40)$$

is a lower-bound of the likelihood parameterized with variational parameters $\Xi \in \mathbb{R}^{L \times M}$.

The condition (9.35) allows us to form a parametric lower-bound in the (unnormalized) isotropic Gaussian form, which we derive as follows. Any function $f(x)$ with bounded curvature $\frac{\partial^2 f}{\partial x^2} \leq \kappa$ can be upper-bounded by the following quadratic function:

$$f(x) \leq \kappa(x - \xi)^2 + \left. \frac{\partial f}{\partial x} \right|_{x=\xi} (x - \xi) + f(\xi) \quad \text{for any } \xi \in \mathbb{R}. \quad (9.41)$$

Therefore, it holds that, for any $\xi \in \mathbb{R}$,

$$-\log \phi_{l,m}(v, u) \leq \frac{(u - \xi)^2}{2\sigma^2} + g(v, \xi)(u - \xi) - \log \phi_{l,m}(v, \xi), \quad (9.42)$$

where

$$g(v, \xi) = -\left. \frac{\partial \log \phi_{l,m}(v, u)}{\partial u} \right|_{u=\xi}.$$

The left graph in Figure 9.7 shows the parametric quadratic upper-bounds (9.42) for the Bernoulli likelihood with sigmoid parameterization.

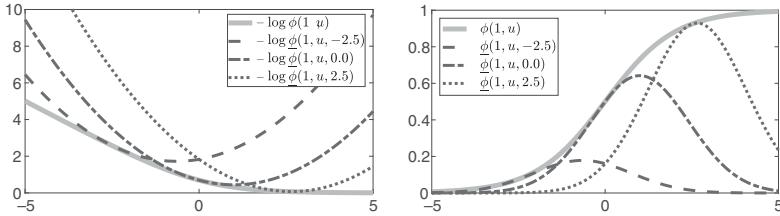


Figure 9.7 Parametric quadratic (Gaussian-form) bounds for the Bernoulli likelihood with sigmoid parameterization, $\phi(v, u) = e^{vu}/(1 + e^u)$, for $v = 1$. Left: the negative log-likelihood (the left-hand side of Eq. (9.42)) and its quadratic upper-bounds (the right-hand side of Eq. (9.42)) for $\xi = -2.5, 0.0, 2.5$. Right: the likelihood function $\phi(1, \xi)$ and its Gaussian-form lower-bounds (9.43) for $\xi = -2.5, 0.0, 2.5$.

Since $\log(\cdot)$ is a monotonic function, we can adopt the following parametric lower-bound of $\phi_{l,m}(v, u)$:

$$\begin{aligned}
\phi_{l,m}(u, \xi) &= \exp\left(-\left(\frac{(u - \xi)^2}{2\sigma^2} + g(v, \xi)(u - \xi) - \log \phi_{l,m}(v, \xi)\right)\right) \\
&= \phi_{l,m}(v, \xi) \exp\left(-\frac{1}{2\sigma^2}((u - \xi)^2 + 2\sigma^2 g(v, \xi)(u - \xi))\right) \\
&= \phi_{l,m}(v, \xi) \exp\left(-\frac{1}{2\sigma^2}\left((u - \xi + \sigma^2 g(v, \xi))^2 - (\sigma^2 g(v, \xi))^2\right)\right) \\
&= \phi_{l,m}(v, \xi) \exp\left(\frac{\sigma^2}{2}g^2(v, \xi)\right) \exp\left(-\frac{1}{2\sigma^2}((\xi - \sigma^2 g(v, \xi)) - u)^2\right) \\
&= \sqrt{2\pi\sigma^2} \phi_{l,m}(v, \xi) \exp\left(\frac{\sigma^2}{2}g^2(v, \xi)\right) \text{Gauss}_1\left(\xi - \sigma^2 g(v, \xi); u, \sigma^2\right).
\end{aligned} \tag{9.43}$$

The right graph in Figure 9.7 shows the parametric Gaussian-form lower-bounds (9.43) for the Bernoulli likelihood with sigmoid parameterization.

Substituting Eq. (9.43) into Eq. (9.39) gives

$$\begin{aligned}
\bar{F}(r, \Xi) &= - \sum_{l=1}^L \sum_{m=1}^M \left(\frac{1}{2} \log(2\pi\sigma^2) + \log \phi_{l,m}(V_{l,m}, \Xi_{l,m}) + \frac{\sigma^2}{2} g^2(V_{l,m}, \Xi_{l,m}) \right) \\
&\quad + \left\langle \log \frac{r_A(A)r_B(B)}{\prod_{l=1}^L \prod_{m=1}^M \text{Gauss}_1(\Xi_{l,m} - \sigma^2 g(V_{l,m}, \Xi_{l,m}); \tilde{\mathbf{a}}_l^\top \tilde{\mathbf{a}}_m, \sigma^2)} p(A)p(B) \right\rangle_{r_A(A)r_B(B)} \\
&= - \sum_{l=1}^L \sum_{m=1}^M \left(\frac{1}{2} \log(2\pi\sigma^2) + \log \phi_{l,m}(V_{l,m}, \Xi_{l,m}) + \frac{\sigma^2}{2} g^2(V_{l,m}, \Xi_{l,m}) \right) \\
&\quad + \left\langle \log \frac{r_A(A)r_B(B)}{(2\pi\sigma^2)^{-LM/2} \exp(-\frac{1}{2\sigma^2} \|\tilde{\mathbf{V}} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2)} p(A)p(B) \right\rangle_{r_A(A)r_B(B)},
\end{aligned} \tag{9.44}$$

where $\tilde{\mathbf{V}} \in \mathbb{R}^{L \times M}$ is a matrix such that

$$\tilde{\mathbf{V}}_{l,m} = \Xi_{l,m} - \sigma^2 g(V_{l,m}, \Xi_{l,m}). \tag{9.45}$$

The first term in Eq. (9.44) does not depend on r and the second term is equal to the free energy of the fully observed isotropic Gaussian MF model with the observed matrix \mathbf{V} replaced with $\check{\mathbf{V}}$. Therefore, given the variational parameter Ξ , we can partially solve the minimization problem (9.38) with respect to r by applying the global VB solver (Algorithm 15). The solution is Gaussian in the following form:

$$r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}), \quad \text{where} \quad (9.46)$$

$$r_A(\mathbf{A}) = \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \mathbf{I}_M \otimes \widehat{\Sigma}_A) \propto \exp\left(-\frac{\text{tr}\left((\mathbf{A} - \widehat{\mathbf{A}})\widehat{\Sigma}_A^{-1}(\mathbf{A} - \widehat{\mathbf{A}})^\top\right)}{2}\right), \quad (9.47)$$

$$r_B(\mathbf{B}) = \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \mathbf{I}_L \otimes \widehat{\Sigma}_B) \propto \exp\left(-\frac{\text{tr}\left((\mathbf{B} - \widehat{\mathbf{B}})\widehat{\Sigma}_B^{-1}(\mathbf{B} - \widehat{\mathbf{B}})^\top\right)}{2}\right). \quad (9.48)$$

Here the mean and the covariance parameters $\widehat{\mathbf{A}}, \widehat{\Sigma}_A, \widehat{\mathbf{B}}, \widehat{\Sigma}_B$ are another set of variational parameters.

Given the optimal r specified by Eq. (9.46), the free energy bound (9.44) is written (as a function of Ξ) as follows:

$$\begin{aligned} \min_r \bar{F}(r, \Xi) &= - \sum_{l=1}^L \sum_{m=1}^M \left(\log \phi_{l,m}(V_{l,m}, \Xi_{l,m}) + \frac{\sigma^2}{2} g^2(V_{l,m}, \Xi_{l,m}) \right) \\ &\quad - \frac{1}{2\sigma^2} \left\langle \|\check{\mathbf{V}} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2 \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} + \text{const.} \\ &= - \sum_{l=1}^L \sum_{m=1}^M \left(\log \phi_{l,m}(V_{l,m}, \Xi_{l,m}) + \frac{\sigma^2}{2} g^2(V_{l,m}, \Xi_{l,m}) \right) \\ &\quad - \frac{1}{2\sigma^2} \|\check{\mathbf{V}} - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top\|_{\text{Fro}}^2 + \text{const.} \\ &= - \sum_{l=1}^L \sum_{m=1}^M \left(\log \phi_{l,m}(V_{l,m}, \Xi_{l,m}) + \frac{\sigma^2}{2} g^2(V_{l,m}, \Xi_{l,m}) \right) \\ &\quad - \frac{1}{2\sigma^2} \sum_{l=1}^L \sum_{m=1}^M \left(\check{V}_{l,m} - \widehat{U}_{l,m} \right)^2 + \text{const.}, \end{aligned} \quad (9.49)$$

where

$$\widehat{\mathbf{U}} = \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top.$$

The second-to-last equation in Eq. (9.43), together with Eq. (9.45), implies that

$$\log \underline{\phi}_{l,m}(\widehat{U}_{l,m}, \Xi_{l,m}) = \log \phi_{l,m}(V_{l,m}, \Xi_{l,m}) + \frac{\sigma^2}{2} g^2(V_{l,m}, \Xi_{l,m}) - \frac{1}{2\sigma^2} (\check{V}_{l,m} - \widehat{U}_{l,m})^2,$$

with which Eq. (9.49) is written as

$$\min_r \bar{F}(r, \Xi) = - \sum_{l=1}^L \sum_{m=1}^M \log \underline{\phi}_{l,m}(\widehat{U}_{l,m}, \Xi_{l,m}) + \text{const.} \quad (9.50)$$

Algorithm 19 Iterative singular value shrinkage algorithm for nonconjugate MF (with missing entries).

- 1: Set the noise variance σ^2 with which the condition (9.35) tightly holds, and initialize the variational parameters to $\Xi = \mathbf{0}_{(L,M)}$.
 - 2: Compute \check{V} by Eq. (9.45).
 - 3: Compute the VB posterior (9.46) by applying the global solver (Algorithm 15) with \check{V} substituted for V .
 - 4: Update Ξ by Eq. (9.51).
 - 5: Iterate Steps 2 through 4 until convergence.
-

Since $\log \underline{\phi}_{l,m}(u, \xi)$ is the quadratic upper-bound (the right-hand side in Eq. (9.42)) of $-\log \phi_{l,m}(v, u)$, which is tight at $u = \widehat{U}_{l,m}$ when $\xi = \widehat{U}_{l,m}$, the minimizer of Eq. (9.50) with respect to Ξ is given by

$$\widehat{\Xi} \equiv \operatorname{argmin}_{\Xi} \min_r \overline{F}(r, \Xi) = \widehat{U}. \quad (9.51)$$

In summary, to solve the joint minimization problem (9.38), we can iteratively update r and Ξ . The update of r can be performed by the global solver (Algorithm 15) with the observed matrix V replaced with \check{V} , defined by Eq. (9.45). The update of Ξ is simply performed by Eq. (9.51). Algorithm 19 summarizes this procedure, where $\mathbf{0}_{(d_1, d_2)}$ denotes the $d_1 \times d_2$ matrix with all entries equal to zero. Seeger and Bouchard (2012) empirically showed that this *iterative singular value shrinkage* algorithm significantly outperforms the MAP solution at comparable computational costs. They also proposed an efficient way to perform SVD when V is huge but sparsely observed, based on the techniques proposed by Tomioka et al. (2010).

9.4.3 Examples of Nonconjugate MF

In this subsection, we introduce a few examples of model likelihood $\phi_{l,m}(v, u)$, which satisfy Condition (9.35), and give the corresponding derivatives of the negative log likelihood.

Isotropic Gaussian MF with Missing Entries

If we let

$$\phi_{l,m}(v, u) = \begin{cases} \text{Gauss}_1(v; u, \sigma^2) & \text{if } (l, m) \in \Lambda, \\ 1 & \text{otherwise,} \end{cases} \quad (9.52)$$

where Λ denotes the set of observed entries, the model distribution (9.32) corresponds to the model distribution (3.44) of MF with missing entries. The first and the second derivatives of the negative log likelihood are given as follows:

$$\begin{aligned} -\frac{\partial \log \phi_{l,m}(v, u)}{\partial u} &= \begin{cases} \frac{1}{\sigma^2} (u - v) & \text{if } (l, m) \in \Lambda, \\ 0 & \text{otherwise,} \end{cases} \\ -\frac{\partial^2 \log \phi_{l,m}(v, u)}{\partial u^2} &= \begin{cases} \frac{1}{\sigma^2} & \text{if } (l, m) \in \Lambda, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (9.53)$$

Bernoulli MF with Sigmoid Parameterization

The *Bernoulli distribution* with *sigmoid* parameterization is suitable for binary observations, i.e., $V \in \{0, 1\}^{L \times M}$:

$$\phi_{l,m}(v, u) = \begin{cases} \frac{e^{vu}}{1 + e^u} & \text{if } (l, m) \in \Lambda, \\ 1 & \text{otherwise.} \end{cases} \quad (9.54)$$

The first and the second derivatives are given as follows:

$$\begin{aligned} -\frac{\partial \log \phi_{l,m}(v, u)}{\partial u} &= \begin{cases} \frac{1}{1+e^{-u}} - v & \text{if } (l, m) \in \Lambda, \\ 0 & \text{otherwise,} \end{cases} \\ -\frac{\partial^2 \log \phi_{l,m}(v, u)}{\partial u^2} &= \begin{cases} \frac{1}{(1+e^{-u})(1+e^u)} & \text{if } (l, m) \in \Lambda, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (9.55)$$

It holds that

$$-\frac{\partial^2 \log \phi_{l,m}(v, u)}{\partial u^2} \leq \frac{1}{4},$$

and therefore, the noise variance should be set to $\sigma^2 = 4$, which satisfies the condition (9.35). Figure 9.7 was depicted for this model.

Poisson MF

The *Poisson distribution* is suitable for count data, i.e., $V \in \{0, 1, 2, \dots\}^{L \times M}$:

$$\phi_{l,m}(v, u) = \begin{cases} \lambda^v(u) e^{-\lambda(u)} & \text{if } (l, m) \in \Lambda, \\ 1 & \text{otherwise,} \end{cases} \quad (9.56)$$

where $\lambda(u)$ is the *link function*. Since a common choice $\lambda(u) = e^u$ for the link function gives unbounded curvature for large u , Seeger and Bouchard (2012)

proposed to use another link function $\lambda(u) = \log(1 + e^u)$. The first derivative is given as follows:

$$-\frac{\partial \log \phi_{l,m}(v, u)}{\partial u} = \begin{cases} \frac{1}{1+e^{-u}} \left(1 - \frac{v}{\lambda(u)}\right) & \text{if } (l, m) \in \Lambda, \\ 0 & \text{otherwise.} \end{cases} \quad (9.57)$$

It was confirmed that the second derivative is upper-bounded as

$$-\frac{\partial^2 \log \phi_{l,m}(v, u)}{\partial u^2} \leq \frac{1}{4} + 0.17v,$$

and therefore, the noise variance should be set to

$$\sigma^2 = \frac{1}{1/4 + 0.17 \max_{l,m} V_{l,m}}.$$

Since the bound can be loose if some of the entries $V_{l,m}$ of the observed matrix are huge compared to the others, overly large counts should be clipped.

10

Global Solver for Low-Rank Subspace Clustering

The nonasymptotic theory, described in Chapter 6, for fully observed matrix factorization (MF) has been extended to other bilinear models. In this chapter, we introduce exact and approximate global variational Bayesian (VB) solvers (Nakajima et al., 2013c) for low-rank subspace clustering (LRSC).

10.1 Problem Description

The LRSC model, introduced in Section 3.4, is defined as

$$p(\mathbf{V}|\mathbf{A}', \mathbf{B}') \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{V} - \mathbf{V}\mathbf{B}'\mathbf{A}'^\top\|_{\text{Fro}}^2\right), \quad (10.1)$$

$$p(\mathbf{A}') \propto \exp\left(-\frac{1}{2} \text{tr}(\mathbf{A}'\mathbf{C}_A^{-1}\mathbf{A}'^\top)\right), \quad (10.2)$$

$$p(\mathbf{B}') \propto \exp\left(-\frac{1}{2} \text{tr}(\mathbf{B}'\mathbf{C}_B^{-1}\mathbf{B}'^\top)\right), \quad (10.3)$$

where $\mathbf{V} \in \mathbb{R}^{L \times M}$ is an observation matrix, and $\mathbf{A}' \in \mathbb{R}^{M \times H}$ and $\mathbf{B}' \in \mathbb{R}^{M \times H}$ for $H \leq \min(L, M)$ are the parameters to be estimated. Note that in this chapter we denote the *original* parameters \mathbf{A}' and \mathbf{B}' with *primes* for convenience. We assume that hyperparameters

$$\mathbf{C}_A = \mathbf{Diag}(c_{a_1}^2, \dots, c_{a_H}^2), \quad \mathbf{C}_B = \mathbf{Diag}(c_{b_1}^2, \dots, c_{b_H}^2),$$

are diagonal and positive definite. The LRSC model is similar to MF. The only difference is that the product $\mathbf{B}'\mathbf{A}'^\top$ of the parameters is further multiplied by \mathbf{V} in Eq. (10.1). Accordingly, we can hope that similar analysis could be applied to LRSC, providing a global solver for LRSC.

We first transform the parameters as

$$\mathbf{A} \leftarrow \boldsymbol{\Omega}_V^{\text{right}\top} \mathbf{A}', \quad \mathbf{B} \leftarrow \boldsymbol{\Omega}_V^{\text{right}\top} \mathbf{B}', \quad \text{where} \quad \mathbf{V} = \boldsymbol{\Omega}_V^{\text{left}} \boldsymbol{\Gamma}_V \boldsymbol{\Omega}_V^{\text{right}\top} \quad (10.4)$$

is the singular value decomposition (SVD) of \mathbf{V} . Here, $\boldsymbol{\Omega}_V^{\text{left}} \in \mathbb{R}^{L \times L}$ and $\boldsymbol{\Omega}_V^{\text{right}} \in \mathbb{R}^{M \times M}$ are orthogonal matrices, and $\boldsymbol{\Gamma}_V \in \mathbb{R}^{L \times M}$ is a (possibly nonsquare) diagonal matrix with nonnegative diagonal entries aligned in nonincreasing order, i.e., $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_{\min(L,M)}$. After this transformation, the LRSC model (10.1) through (10.3) is rewritten as

$$p(\boldsymbol{\Gamma}_V | \mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\boldsymbol{\Gamma}_V - \boldsymbol{\Gamma}_V \mathbf{B} \mathbf{A}^\top\|_{\text{Fro}}^2\right), \quad (10.5)$$

$$p(\mathbf{A}) \propto \exp\left(-\frac{1}{2} \text{tr}(\mathbf{A} \mathbf{C}_A^{-1} \mathbf{A}^\top)\right), \quad (10.6)$$

$$p(\mathbf{B}) \propto \exp\left(-\frac{1}{2} \text{tr}(\mathbf{B} \mathbf{C}_B^{-1} \mathbf{B}^\top)\right). \quad (10.7)$$

The transformation (10.4) does not affect much the derivation of the VB learning algorithm. The following summarizes the result obtained in Section 3.4 with the transformed parameters \mathbf{A} and \mathbf{B} . The solution of the VB learning problem,

$$\begin{aligned} \widehat{r} &= \underset{r}{\operatorname{argmin}} F(r) \quad \text{s.t.} \quad r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}), \quad \text{where} \\ F &= \left\langle \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{p(\boldsymbol{\Gamma}_V | \mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}, \end{aligned} \quad (10.8)$$

has the following form:

$$\begin{aligned} r(\mathbf{A}) &\propto \exp\left(-\frac{\text{tr}\left((\mathbf{A} - \widehat{\mathbf{A}})\widehat{\Sigma}_A^{-1}(\mathbf{A} - \widehat{\mathbf{A}})^\top\right)}{2}\right), \\ r(\mathbf{B}) &\propto \exp\left(-\frac{(\check{\mathbf{b}} - \widehat{\mathbf{b}})^\top \widehat{\Sigma}_B^{-1}(\check{\mathbf{b}} - \widehat{\mathbf{b}})}{2}\right), \end{aligned} \quad (10.9)$$

for $\check{\mathbf{b}} = \text{vec}(\mathbf{B}) \in \mathbb{R}^{MH}$, and the free energy can be explicitly written as

$$\begin{aligned} 2F &= LM \log(2\pi\sigma^2) + \frac{\|\boldsymbol{\Gamma}_V - \boldsymbol{\Gamma}_V \widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top\|_{\text{Fro}}^2}{\sigma^2} + M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + \log \frac{\det(\mathbf{C}_B \otimes \mathbf{I}_M)}{\det(\widehat{\Sigma}_B)} \\ &\quad - 2MH + \text{tr} \left\{ \mathbf{C}_A^{-1} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) \right\} + \text{tr} \left\{ \mathbf{C}_B^{-1} \widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} \right\} + \text{tr} \left\{ (\mathbf{C}_B^{-1} \otimes \mathbf{I}_M) \widehat{\Sigma}_B \right\} \\ &\quad + \text{tr} \left\{ \sigma^{-2} \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \left(-\widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\mathbf{B}}^\top + \left\langle \mathbf{B}(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A) \mathbf{B}^\top \right\rangle_{r(\mathbf{B})} \right) \right\}. \end{aligned} \quad (10.10)$$

Therefore, the variational parameters $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B)$ can be obtained by solving the following problem:

$$\text{Given } \mathbf{C}_A, \mathbf{C}_B \in \mathbb{D}_{++}^H, \sigma^2 \in \mathbb{R}_{++}, \min_{(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B)} F, \quad (10.11)$$

$$\text{s.t. } \widehat{\mathbf{A}}, \widehat{\mathbf{B}} \in \mathbb{R}^{M \times H}, \widehat{\Sigma}_A \in \mathbb{S}_{++}^H, \widehat{\Sigma}_B \in \mathbb{S}_{++}^{MH}. \quad (10.12)$$

The stationary conditions with respect to the variational parameters are given by

$$\widehat{\mathbf{A}} = \frac{1}{\sigma^2} \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \widehat{\mathbf{B}} \widehat{\Sigma}_A, \quad (10.13)$$

$$\widehat{\Sigma}_A = \sigma^2 \left(\left\langle \mathbf{B}^\top \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \mathbf{B} \right\rangle_{r(\mathbf{B})} + \sigma^2 \mathbf{C}_A^{-1} \right)^{-1}, \quad (10.14)$$

$$\widehat{\mathbf{b}} = \frac{\widehat{\Sigma}_B}{\sigma^2} \text{vec}(\boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \widehat{\mathbf{A}}), \quad (10.15)$$

$$\widehat{\Sigma}_B = \sigma^2 \left((\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A) \otimes \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V + \sigma^2 (\mathbf{C}_B^{-1} \otimes \mathbf{I}_M) \right)^{-1}. \quad (10.16)$$

For empirical VB (EVB) learning, we solve the problem,

Given $\sigma^2 \in \mathbb{R}_{++}$,

$$\min_{(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B, \mathbf{C}_A, \mathbf{C}_B)} F \quad (10.17)$$

$$\text{subject to } \widehat{\mathbf{A}}, \widehat{\mathbf{B}} \in \mathbb{R}^{M \times H}, \widehat{\Sigma}_A \in \mathbb{S}_{++}^H, \widehat{\Sigma}_B \in \mathbb{S}_{++}^{MH}, \mathbf{C}_A, \mathbf{C}_B \in \mathbb{D}_{++}^H, \quad (10.18)$$

for which the stationary conditions with respect to the hyperparameters are given by

$$c_{a_h}^2 = \|\widehat{\mathbf{a}}_h\|^2 / M + (\widehat{\Sigma}_A)_{h,h}, \quad (10.19)$$

$$c_{b_h}^2 = \left(\|\widehat{\mathbf{b}}_h\|^2 + \text{tr}(\widehat{\Sigma}_B^{(h,h)}) \right) / M, \quad (10.20)$$

$$\widehat{\sigma}^2 = \frac{\text{tr} \left(\boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \left(\mathbf{I}_M - 2 \widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top + \left\langle \mathbf{B} (\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A) \mathbf{B}^\top \right\rangle_{r(\mathbf{B})} \right) \right)}{LM}. \quad (10.21)$$

In deriving the global VB solution of fully observed MF in Chapter 6, the following two facts were essential. First, a large portion of the degrees of freedom of the *original* variational parameters are irrelevant (see Section 6.3), and the optimization problem can be decomposed into subproblems, each of which has only a small number of unknown variables. Second, the stationary conditions of each subproblem is written as a *polynomial system* (a set of

polynomial equations). These two facts also apply to the LRSC model, which allows us to derive an *exact* global VB solver (EGVBS). However, each of the decomposed subproblems still has too many unknowns whose number is proportional to the problem size, and therefore EGVBS is still computationally demanding for typical problem sizes. As an alternative, we also derive an approximate global VB solver (AGVBS) by imposing an additional constraint, which allows further decomposition of the problem into subproblems with a constant number of unknowns.

In this chapter, we first find irrelevant degrees of freedom of the variational parameters and decompose the VB learning problem. Then we derive EGVBS and AGVBS and empirically show their usefulness.

10.2 Conditions for VB Solutions

Let J ($\leq \min(L, M)$) be the rank of the observed matrix \mathbf{V} . For simplicity, we assume that no pair of positive singular values of \mathbf{V} coincide with each other, i.e.,

$$\gamma_1 > \gamma_2 > \cdots > \gamma_J > 0.$$

This holds with probability 1 if \mathbf{V} is contaminated with Gaussian noise, as the LRSC model (10.1) assumes. Since $(\boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V)_{m,m'}$ is zero for $m > J$ or $m' > J$, Eqs. (10.13) and (10.15) imply that

$$\widehat{A}_{m,h} = \widehat{B}_{m,h} = 0 \quad \text{for } m > J. \quad (10.22)$$

Similarly to Lemma 6.1 for the fully observed MF, we can prove the following lemma:

Lemma 10.1 *Any local solution of the problem (10.11) is a stationary point of the free energy (10.10).*

Proof Since

$$\left\| \boldsymbol{\Gamma}_V - \boldsymbol{\Gamma}_V \widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top \right\|_{\text{Fro}}^2 \geq 0,$$

and

$$\begin{aligned} \text{tr} \left\{ \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \left(-\widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\mathbf{B}}^\top + \left\langle \mathbf{B} (\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A) \mathbf{B}^\top \right\rangle_{r(\mathbf{B})} \right) \right\} \\ = M \cdot \text{tr} \left\{ \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \left\langle \mathbf{B} \widehat{\Sigma}_A \mathbf{B}^\top \right\rangle_{r(\mathbf{B})} \right\} \geq 0, \end{aligned}$$

the free energy (10.10) is lower-bounded as

$$\begin{aligned} 2F \geq & -M \log \det(\widehat{\Sigma}_A) - \log \det(\widehat{\Sigma}_B) \\ & + \text{tr} \left\{ C_A^{-1} \left(\widehat{A}^\top \widehat{A} + M \widehat{\Sigma}_A \right) \right\} + \text{tr} \left\{ C_B^{-1} \widehat{B}^\top \widehat{B} \right\} + \text{tr} \left\{ (C_B^{-1} \otimes I_M) \widehat{\Sigma}_B \right\} + \tau, \end{aligned} \quad (10.23)$$

where τ is a finite constant. The right-hand side of Eq. (10.23) diverges to $+\infty$ if any entry of \widehat{A} or \widehat{B} goes to $+\infty$ or $-\infty$. Also it diverges if any eigenvalue of $\widehat{\Sigma}_A$ or $\widehat{\Sigma}_B$ goes to $+0$ or ∞ . This implies that no local solution exists on the boundary of (the closure of) the domain (10.12). Since the free energy is differentiable in the domain (10.12), any local minimizer is a stationary point.

For any (diagonalized) observed matrix Γ_V , the free energy (10.10) can be finite, for example, at $\widehat{A} = \mathbf{0}_{M,H}$, $\widehat{B} = \mathbf{0}_{M,H}$, $\widehat{\Sigma}_A = I_H$, and $\widehat{\Sigma}_B = I_{MH}$. Therefore, at least one minimizer always exists, which completes the proof of Lemma 10.1. \square

Lemma 10.1 implies that Eqs. (10.13) through (10.16) hold at any local solution.

10.3 Irrelevant Degrees of Freedom

Also similarly to Theorem 6.4, we have the following theorem:

Theorem 10.2 *When $C_A C_B$ is nondegenerate (i.e., $c_{ah} c_{bh} > c_{ah'} c_{bh'}$ for any pair $h < h'$), $(\widehat{A}, \widehat{B}, \widehat{\Sigma}_A, \widehat{\Sigma}_B)$ are diagonal for any solution of the problem (10.11). When $C_A C_B$ is degenerate, any solution has an equivalent solution with diagonal $(\widehat{A}, \widehat{B}, \widehat{\Sigma}_A, \widehat{\Sigma}_B)$.*

Theorem 10.2 significantly reduces the complexity of the optimization problem, and furthermore makes the problem separable, as seen in Section 10.5.

10.4 Proof of Theorem 10.2

Similarly to Section 6.4, we separately consider the following three cases:

Case 1 When no pair of diagonal entries of $C_A C_B$ coincide.

Case 2 When all diagonal entries of $C_A C_B$ coincide.

Case 3 When (not all but) some pairs of diagonal entries of $C_A C_B$ coincide.

10.4.1 Diagonality Implied by Optimality

We can prove the following lemma, which is an extension of Lemma 6.2.

Lemma 10.3 *Let $\Gamma, \Omega, \Phi \in \mathbb{R}^{H \times H}$ be a nondegenerate diagonal matrix, an orthogonal matrix, and a symmetric matrix, respectively. Let $\{\Lambda^{(k)}, \Lambda'^{(k)} \in \mathbb{R}^{H \times H}; k = 1, \dots, K\}$ be arbitrary diagonal matrices, and $\{\Psi^{(k')} \in \mathbb{R}^{H \times H}; k' = 1, \dots, K'\}$ be arbitrary symmetric matrices. If*

$$G(\Omega) = \text{tr} \left\{ \Gamma \Omega \Phi \Omega^\top + \sum_{k=1}^K \Lambda^{(k)} \Omega \Lambda'^{(k)} \Omega^\top + \sum_{k'=1}^{K'} \Omega \Psi^{(k')} \right\} \quad (10.24)$$

is minimized or maximized (as a function of Ω , given $\Gamma, \Phi, \{\Lambda^{(k)}, \Lambda'^{(k)}\}, \{\Psi^{(k')}\}$) when $\Omega = I_H$, then Φ is diagonal. Here, K and K' can be any natural numbers including $K = 0$ and $K' = 0$ (when the second and the third terms, respectively, do not exist).

Proof Let

$$\Phi = \Omega' \Gamma' \Omega'^\top \quad (10.25)$$

be the eigenvalue decomposition of Φ . Let $\gamma, \gamma', \{\lambda^{(k)}\}, \{\lambda'^{(k)}\}$ be the vectors consisting of the diagonal entries of $\Gamma, \Gamma', \{\Lambda^{(k)}\}, \{\Lambda'^{(k)}\}$, respectively, i.e.,

$$\Gamma = \text{Diag}(\gamma), \quad \Gamma' = \text{Diag}(\gamma'), \quad \Lambda^{(k)} = \text{Diag}(\lambda^{(k)}), \quad \Lambda'^{(k)} = \text{Diag}(\lambda'^{(k)}).$$

Then, Eq. (10.24) can be written as

$$\begin{aligned} G(\Omega) &= \text{tr} \left\{ \Gamma \Omega \Phi \Omega^\top + \sum_{k=1}^K \Lambda^{(k)} \Omega \Lambda'^{(k)} \Omega^\top + \sum_{k'=1}^{K'} \Omega \Psi^{(k')} \right\} \\ &= \gamma^\top Q \gamma' + \sum_{k=1}^K \lambda^{(k)\top} R \lambda'^{(k)} + \sum_{k'=1}^{K'} \text{tr} \{ \Omega \Psi^{(k')} \}, \end{aligned} \quad (10.26)$$

where

$$Q = (\Omega \Omega') \odot (\Omega \Omega'), \quad R = \Omega \odot \Omega.$$

Here, \odot denotes the Hadamard product.

Using this expression, we will prove that Φ is diagonal if $\Omega = I_H$ minimizes or maximizes Eq. (10.26). Let us consider a bilateral perturbation $\Omega = A$ such that the 2×2 matrix $A_{(h,h')}$ for $h \neq h'$ consisting of the h th and the h' th columns and rows form a 2×2 orthogonal matrix,

$$A_{(h,h')} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix},$$

and the remaining entries coincide with those of the identity matrix. Then, the elements of \mathbf{Q} become

$$Q_{i,j} = \begin{cases} (\Omega'_{h,j} \cos \theta - \Omega'_{h',j} \sin \theta)^2 & \text{if } i = h, \\ (\Omega'_{h,j} \sin \theta + \Omega'_{h',j} \cos \theta)^2 & \text{if } i = h', \\ \Omega'^2_{i,j} & \text{otherwise,} \end{cases}$$

and Eq. (10.26) can be written as a function of θ as follows:

$$\begin{aligned} G(\theta) = & \sum_{j=1}^H \left\{ \gamma_h (\Omega'_{h,j} \cos \theta - \Omega'_{h',j} \sin \theta)^2 + \gamma_{h'} (\Omega'_{h,j} \sin \theta + \Omega'_{h',j} \cos \theta)^2 \right\} \gamma'_j \\ & + \sum_{k=1}^K \begin{pmatrix} \lambda_h^{(k')} & \lambda_{h'}^{(k')} \end{pmatrix} \begin{pmatrix} \cos^2 \theta & \sin^2 \theta \\ \sin^2 \theta & \cos^2 \theta \end{pmatrix} \begin{pmatrix} \lambda_h^{(k')} \\ \lambda_{h'}^{(k')} \end{pmatrix} \\ & + \sum_{k'=1}^{K'} \left(\Psi_{h,h}^{(k')} \cos \theta - \Psi_{h',h}^{(k')} \sin \theta + \Psi_{h,h'}^{(k')} \sin \theta + \Psi_{h',h'}^{(k')} \cos \theta \right) + \text{const.} \end{aligned} \quad (10.27)$$

Since Eq. (10.27) is differentiable at $\theta = 0$, our assumption that Eq. (10.26) is minimized or maximized when $\mathbf{Q} = \mathbf{I}_H$ requires that $\theta = 0$ is a stationary point of Eq. (10.27) for any $h \neq h'$. Therefore, it holds that

$$\begin{aligned} 0 = \frac{\partial G}{\partial \theta} \Big|_{\theta=0} = & \left[2 \sum_j \left\{ \gamma_h (\Omega'_{h,j} \cos \theta - \Omega'_{h',j} \sin \theta) (-\Omega'_{h,j} \sin \theta - \Omega'_{h',j} \cos \theta) \right. \right. \\ & + \gamma_{h'} (\Omega'_{h,j} \sin \theta + \Omega'_{h',j} \cos \theta) (\Omega'_{h,j} \cos \theta - \Omega'_{h',j} \sin \theta) \left. \right\} \gamma'_j \\ & + \sum_{k'=1}^{K'} \left(-\Psi_{h,h}^{(k')} \sin \theta - \Psi_{h',h}^{(k')} \cos \theta + \Psi_{h,h'}^{(k')} \cos \theta - \Psi_{h',h'}^{(k')} \sin \theta \right) \Big] \Big|_{\theta=0} \\ = & 2 (\gamma_{h'} - \gamma_h) \sum_j \Omega'_{h,j} \gamma'_j \Omega'_{h',j} + \sum_{k'=1}^{K'} (\Psi_{h,h'}^{(k')} - \Psi_{h',h}^{(k')}) \\ = & 2 (\gamma_{h'} - \gamma_h) \Phi_{h,h'}. \end{aligned} \quad (10.28)$$

In the last equation, we used Eq. (10.25) and the assumption that $\{\Psi^{(k')}\}$ are symmetric. Since we assume that $\boldsymbol{\Gamma}$ is nondegenerate ($\gamma_h \neq \gamma_{h'}$ for $h \neq h'$), Eq. (10.28) implies that $\boldsymbol{\Phi}$ is diagonal, which completes the proof of Lemma 10.3. \square

10.4.2 Proof for Case 1

Assume that $(\mathbf{A}^*, \mathbf{B}^*, \boldsymbol{\Sigma}_A^*, \check{\boldsymbol{\Sigma}}_B^*)$ is a minimizer, and consider the following variation defined with an arbitrary $H \times H$ orthogonal matrix \mathbf{Q}_1 :

$$\widehat{\mathbf{A}} = \mathbf{A}^* \mathbf{C}_B^{1/2} \mathbf{Q}_1^\top \mathbf{C}_B^{-1/2}, \quad (10.29)$$

$$\widehat{\mathbf{B}} = \mathbf{B}^* \mathbf{C}_B^{-1/2} \boldsymbol{\Omega}_1^\top \mathbf{C}_B^{1/2}, \quad (10.30)$$

$$\widehat{\Sigma}_A = \mathbf{C}_B^{-1/2} \boldsymbol{\Omega}_1 \mathbf{C}_B^{1/2} \boldsymbol{\Sigma}_A^* \mathbf{C}_B^{1/2} \boldsymbol{\Omega}_1^\top \mathbf{C}_B^{-1/2}, \quad (10.31)$$

$$\widehat{\check{\Sigma}}_B = (\mathbf{C}_B^{1/2} \boldsymbol{\Omega}_1 \mathbf{C}_B^{-1/2} \otimes \mathbf{I}_M) \check{\Sigma}_B^* (\mathbf{C}_B^{-1/2} \boldsymbol{\Omega}_1^\top \mathbf{C}_B^{1/2} \otimes \mathbf{I}_M). \quad (10.32)$$

Then the free energy (10.10) can be written as a function of $\boldsymbol{\Omega}_1$:

$$2F(\boldsymbol{\Omega}_1) = \text{tr} \left\{ (\mathbf{C}_A^{-1} \mathbf{C}_B^{-1} \boldsymbol{\Omega}_1 \mathbf{C}_B^{1/2} (A^{*\top} A^* + M \boldsymbol{\Sigma}_A^*) \mathbf{C}_B^{1/2} \boldsymbol{\Omega}_1^\top) + \text{const.} \right\} \quad (10.33)$$

Since Eq. (10.33) is minimized when $\boldsymbol{\Omega}_1 = \mathbf{I}_H$ by assumption, Lemma 10.3 implies that

$$\mathbf{C}_B^{1/2} (A^{*\top} A^* + M \boldsymbol{\Sigma}_A^*) \mathbf{C}_B^{1/2}$$

is diagonal. Therefore,

$$\boldsymbol{\Phi}_1 = A^{*\top} A^* + M \boldsymbol{\Sigma}_A^* \quad (10.34)$$

is diagonal, with which Eq. (10.16) implies that $\check{\Sigma}_B^*$ is diagonal.

Since we have proved the diagonality of $\check{\Sigma}_B^*$, the expectations in Eqs. (10.10) and (10.14), respectively, can be expressed in the following simple forms at the solution $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\Sigma}_A, \widehat{\Sigma}_B) = (A^*, \mathbf{B}^*, \boldsymbol{\Sigma}_A^*, \check{\Sigma}_B^*)$:

$$\left\langle \mathbf{B} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\boldsymbol{\Sigma}}_A \right) \mathbf{B}^\top \right\rangle_{r_B(\mathbf{B})} = \widehat{\mathbf{B}} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\boldsymbol{\Sigma}}_A \right) \widehat{\mathbf{B}}^\top + \boldsymbol{\Xi}_{\boldsymbol{\Phi}_1}, \quad (10.35)$$

$$\left\langle \mathbf{B}^\top \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \mathbf{B} \right\rangle_{r_B(\mathbf{B})} = \widehat{\mathbf{B}}^\top \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \widehat{\mathbf{B}} + \boldsymbol{\Xi}_{\boldsymbol{\Gamma}_V}, \quad (10.36)$$

where $\boldsymbol{\Xi}_{\boldsymbol{\Gamma}_V} \in \mathbb{R}^{H \times H}$ and $\boldsymbol{\Xi}_{\boldsymbol{\Phi}_1} \in \mathbb{R}^{M \times M}$ are diagonal matrices with their entries given by

$$\begin{aligned} (\boldsymbol{\Xi}_{\boldsymbol{\Phi}_1})_{m,m} &= \sum_{h=1}^H \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\boldsymbol{\Sigma}}_A \right)_{h,h} \widehat{\sigma}_{B_{m,h}}^2, \\ (\boldsymbol{\Xi}_{\boldsymbol{\Gamma}_V})_{h,h} &= \sum_{m=1}^M \gamma_m^2 \widehat{\sigma}_{B_{m,h}}^2. \end{aligned}$$

Here $\{\widehat{\sigma}_{B_{m,h}}^2\}$ are the diagonal entries of $\widehat{\Sigma}_B$ such that

$$\widehat{\Sigma}_B = \text{Diag}((\widehat{\sigma}_{B_{1,1}}^2, \dots, \widehat{\sigma}_{B_{M,1}}^2), (\widehat{\sigma}_{B_{1,2}}^2, \dots, \widehat{\sigma}_{B_{M,2}}^2), \dots, (\widehat{\sigma}_{B_{1,H}}^2, \dots, \widehat{\sigma}_{B_{M,H}}^2)).$$

Next consider the following variation defined with an $M \times M$ matrix $\boldsymbol{\Omega}_2$ such that the upper-left $J \times J$ submatrix is an arbitrary orthogonal matrix and the other entries are zero:

$$\widehat{\mathbf{A}} = \boldsymbol{\Omega}_2^\top \mathbf{A}^*,$$

$$\widehat{\mathbf{B}} = \boldsymbol{\Omega}_2^\top \mathbf{B}^*.$$

Then, by using Eq. (10.35), the free energy (10.10) is written as

$$2F(\boldsymbol{\Omega}_2) = \frac{1}{\sigma^2} \text{tr} \left\{ \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \boldsymbol{\Omega}_2^\top \left(-2\mathbf{B}^* \mathbf{A}^{*\top} + \mathbf{B}^* (\mathbf{A}^{*\top} \mathbf{A}^* + M\widehat{\boldsymbol{\Sigma}}_A) \mathbf{B}^{*\top} \right) \boldsymbol{\Omega}_2 \right\} + \text{const.} \quad (10.37)$$

Applying Lemma 10.3 to the upper-left $J \times J$ submatrix in the trace, and then using Eq. (10.22), we find that

$$\boldsymbol{\Phi}_2 = -2\mathbf{B}^* \mathbf{A}^{*\top} + \mathbf{B}^* (\mathbf{A}^{*\top} \mathbf{A}^* + M\widehat{\boldsymbol{\Sigma}}_A) \mathbf{B}^{*\top} \quad (10.38)$$

is diagonal. Eq. (10.38) also implies that $\mathbf{B}^* \mathbf{A}^{*\top}$ is symmetric.

Consider the following variation defined with an $M \times M$ matrix $\boldsymbol{\Omega}_3$ such that the upper-left $J \times J$ submatrix is an arbitrary orthogonal matrix and the other entries are zero:

$$\widehat{\mathbf{B}} = \boldsymbol{\Omega}_3^\top \mathbf{B}^*.$$

Then the free energy is written as

$$2F(\boldsymbol{\Omega}_3) = \frac{1}{\sigma^2} \text{tr} \left\{ \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \boldsymbol{\Omega}_3^\top \left(-2\mathbf{B}^* \mathbf{A}^{*\top} \right) \right. \\ \left. + \boldsymbol{\Gamma}_V^\top \boldsymbol{\Gamma}_V \boldsymbol{\Omega}_3^\top \left(\mathbf{B}^* (\mathbf{A}^{*\top} \mathbf{A}^* + M\widehat{\boldsymbol{\Sigma}}_A) \mathbf{B}^{*\top} \right) \boldsymbol{\Omega}_3 \right\} + \text{const.} \quad (10.39)$$

Applying Lemma 10.3 to the upper-left $J \times J$ submatrix in the trace, we find that

$$\boldsymbol{\Phi}_3 = \mathbf{B}^* (\mathbf{A}^{*\top} \mathbf{A}^* + M\widehat{\boldsymbol{\Sigma}}_A) \mathbf{B}^{*\top} \quad (10.40)$$

is diagonal. Since Eqs. (10.34) and (10.40) are diagonal, \mathbf{B}^* is diagonal. Consequently, Eq. (10.14) combined with Eq. (10.36) implies that \mathbf{A}^* and $\boldsymbol{\Sigma}_A^*$ are diagonal.

Thus we proved that the solution for $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\Sigma}}_A, \widehat{\boldsymbol{\Sigma}}_B)$ are diagonal, provided that $\mathbf{C}_A \mathbf{C}_B$ is nondegenerate.

10.4.3 Proof for Case 2

When $\mathbf{C}_A \mathbf{C}_B$ is degenerate, there are multiple *equivalent* solutions giving the same free energy (10.10) and the output $\widehat{\mathbf{B}} \mathbf{A}^{*\top}$. In the following, we show that one of the equivalent solutions has diagonal $(\mathbf{A}^*, \mathbf{B}^*, \boldsymbol{\Sigma}_A^*, \boldsymbol{\Sigma}_B^*)$.

Assume that $\mathbf{C}_A \mathbf{C}_B = c^2 \mathbf{I}_H$ for some $c^2 \in \mathbb{R}_{++}$. In this case, the free energy (10.10) is invariant with respect to $\boldsymbol{\Omega}_1$ under the transformation (10.29) through (10.32). Let us focus on the solution with diagonal $\widehat{\boldsymbol{\Sigma}}_B$, which can be obtained by the transform (10.29) through (10.32) with a certain $\boldsymbol{\Omega}_1$ from any solution satisfying Eq. (10.16). Then we can show, in the same way as in the

nondegenerate case, that Eqs. (10.34), (10.38), and (10.40) are diagonal. This proves the existence of a solution such that $(A^*, B^*, \Sigma_A^*, \check{\Sigma}_B^*)$ are diagonal.

10.4.4 Proof for Case 3

When $c_{ah}c_{bh} = c_{ah'}c_{bh'}$ for (not all but) some pairs $h \neq h'$, we can show that $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$ are block diagonal where the blocks correspond to the groups sharing the same $c_{ah}c_{bh}$. In each block, multiple equivalent solutions exist, one of which is a solution such that $(A^*, B^*, \Sigma_A^*, \check{\Sigma}_B^*)$ are diagonal.

This completes the proof of Theorem 10.2. \square

10.5 Exact Global VB Solver (EGVBS)

Theorem 10.2 allows us to focus on the solutions such that $(\widehat{A}, \widehat{B}, \widehat{\Sigma}_A, \widehat{\Sigma}_B)$ are diagonal. Accordingly, we express the solution of the VB learning problem (10.11) with diagonal entries, i.e.,

$$\widehat{A} = \text{Diag}_{M,H}(\widehat{a}_1, \dots, \widehat{a}_H), \quad (10.41)$$

$$\widehat{B} = \text{Diag}_{M,H}(\widehat{b}_1, \dots, \widehat{b}_H), \quad (10.42)$$

$$\widehat{\Sigma}_A = \text{Diag}(\widehat{\sigma}_{a_1}^2, \dots, \widehat{\sigma}_{a_H}^2), \quad (10.43)$$

$$\widehat{\Sigma}_B = \text{Diag}((\widehat{\sigma}_{B_{1,1}}^2, \dots, \widehat{\sigma}_{B_{M,1}}^2), (\widehat{\sigma}_{B_{1,2}}^2, \dots, \widehat{\sigma}_{B_{M,2}}^2), \dots, (\widehat{\sigma}_{B_{1,H}}^2, \dots, \widehat{\sigma}_{B_{M,H}}^2)), \quad (10.44)$$

where $\text{Diag}_{D_1, D_2}(\cdot)$ denotes the $D_1 \times D_2$ diagonal matrix with the specified diagonal entries. Remember that J ($\leq \min(L, M)$) is the rank of the observed matrix V , and $\{\gamma_m\}$ are the singular values arranged in nonincreasing order. Without loss of generality, we assume that $\widehat{a}_h, \widehat{b}_h \in \mathbb{R}_+$ for all $h = 1, \dots, H$.

We can easily obtain the following theorem:

Theorem 10.4 *Any local solution of the VB learning problem (10.11) satisfies, for all $h = 1, \dots, H$,*

$$\widehat{a}_h = \frac{\gamma_h^2}{\sigma^2} \widehat{b}_h \widehat{\sigma}_{a_h}^2, \quad (10.45)$$

$$\widehat{\sigma}_{a_h}^2 = \sigma^2 \left(\gamma_h^2 \widehat{b}_h^2 + \sum_{m=1}^J \gamma_m^2 \widehat{\sigma}_{B_{m,h}}^2 + \frac{\sigma^2}{c_{a_h}^2} \right)^{-1}, \quad (10.46)$$

$$\widehat{b}_h = \frac{\gamma_h^2}{\sigma^2} \widehat{a}_h \widehat{\sigma}_{B_{h,h}}^2, \quad (10.47)$$

$$\widehat{\sigma}_{B_{m,h}}^2 = \begin{cases} \sigma^2 \left(\gamma_m^2 (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2) + \frac{\sigma^2}{c_{b_h}^2} \right)^{-1} & \text{(for } m = 1, \dots, J), \\ c_{b_h}^2 & \text{(for } m = J+1, \dots, M), \end{cases} \quad (10.48)$$

and has the free energy given by

$$2F = LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^J \gamma_h^2}{\sigma^2} + \sum_{h=1}^H 2F_h, \quad \text{where} \quad (10.49)$$

$$\begin{aligned} 2F_h &= M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + \sum_{m=1}^J \log \frac{c_{b_h}^2}{\widehat{\sigma}_{B_{m,h}}^2} - (M+J) + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + \sum_{m=1}^J \widehat{\sigma}_{B_{m,h}}^2}{c_{b_h}^2} \\ &\quad + \frac{1}{\sigma^2} \left\{ \gamma_h^2 (-2\widehat{a}_h \widehat{b}_h + \widehat{b}_h^2 (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2)) + \sum_{m=1}^J \gamma_m^2 \widehat{\sigma}_{B_{m,h}}^2 (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2) \right\}. \end{aligned} \quad (10.50)$$

Proof By substituting the diagonal expression, Eqs. (10.41) through (10.44), into the free energy (10.10), we have

$$\begin{aligned} 2F &= LM \log(2\pi\sigma^2) + M \sum_{h=1}^H \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + \sum_{m=1}^M \sum_{h=1}^H \log \frac{c_{b_h}^2}{\widehat{\sigma}_{B_{m,h}}^2} + \frac{\sum_{h=1}^M \gamma_h^2}{\sigma^2} - 2MH \\ &\quad + \sum_{h=1}^H \left\{ \frac{1}{c_{a_h}^2} (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2) + \frac{1}{c_{b_h}^2} \left(\widehat{b}_h^2 + \sum_{m=1}^M \widehat{\sigma}_{B_{m,h}}^2 \right) \right\} \\ &\quad + \frac{1}{\sigma^2} \sum_{h=1}^H \left\{ \gamma_h^2 (-2\widehat{a}_h \widehat{b}_h + \widehat{b}_h^2 (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2)) + \sum_{m=1}^J \gamma_m^2 \widehat{\sigma}_{B_{m,h}}^2 (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2) \right\}. \end{aligned} \quad (10.51)$$

Eqs. (10.45) through (10.48) are obtained as the stationary conditions of Eq. (10.51) that any solution satisfies, according to Lemma 10.1. By substituting Eq. (10.48) for $m = J+1, \dots, M$ into Eq. (10.51), we obtain Eq. (10.49). \square

For EVB learning, where the prior covariances $\mathbf{C}_A, \mathbf{C}_B$ are also estimated, we have the following theorem:

Theorem 10.5 Any local solution of the EVB learning problem (10.17) satisfies the following. For each $h = 1, \dots, H$, $(\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \{\widehat{\sigma}_{B_{m,h}}^2\}_{m=1}^M, c_{a_h}^2, c_{b_h}^2)$ is either a (positive) stationary point that satisfies Eqs. (10.45) through (10.48) and

$$c_{a_h}^2 = \widehat{a}_h^2 / M + \widehat{\sigma}_{a_h}^2, \quad (10.52)$$

$$c_{b_h}^2 = \left(\widehat{b}_h^2 + \sum_{m=1}^J \widehat{\sigma}_{B_{m,h}}^2 \right) / J, \quad (10.53)$$

or the null local solution defined by

$$\widehat{a}_h = \widehat{b}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = c_{a_h}^2 \rightarrow +0, \quad \widehat{\sigma}_{B_{m,h}}^2 = c_{b_h}^2 \rightarrow +0 \quad (\text{for } m = 1, \dots, M), \quad (10.54)$$

of which the contribution (10.50) to the free energy is

$$F_h \rightarrow +0. \quad (10.55)$$

The total free energy is given by Eq. (10.49).

Proof Considering the derivatives of Eq. (10.51) with respect to $c_{a_h}^2$ and $c_{b_h}^2$, we have

$$2Mc_{a_h}^2 = \widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2, \quad (10.56)$$

$$2Mc_{b_h}^2 = \widehat{b}_h^2 + \sum_{m=1}^M \widehat{\sigma}_{B_{m,h}}^2, \quad (10.57)$$

as stationary conditions. By using Eq. (10.48), we can easily obtain Eqs. (10.52) and (10.53).

Unlike in VB learning, where Lemma 10.1 guarantees that any local solution is a stationary point, there exist nonstationary local solutions in EVB learning. We can confirm that, along any path such that

$$\begin{aligned} \widehat{a}_h, \widehat{b}_h &= 0, \quad \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{B_{m,h}}^2, c_{a_h}^2, c_{b_h}^2 \rightarrow +0 \\ \text{with } \beta_a &= \frac{\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} \text{ and } \beta_b = \frac{\widehat{\sigma}_{B_{m,h}}^2}{c_{b_h}^2} \text{ kept constant,} \end{aligned} \quad (10.58)$$

the free energy contribution (10.50) from the h th component decreases monotonically. Among the possible paths, $\beta_a = \beta_b = 1$ gives the lowest free energy (10.55). \square

Based on Theorem 10.5, we can obtain the following corollary for the global solution.

Corollary 10.6 *The global solution of the EVB learning problem (10.17) can be found in the following way. For each $h = 1, \dots, H$, find all stationary points that satisfy Eqs. (10.45) through (10.48), (10.52), and (10.53), and choose the one giving the minimum free energy contribution F_h . The chosen stationary point is the global solution if $F_h < 0$. Otherwise (including the case where no stationary point exists), the null local solution (10.54) is global.*

Proof For each $h = 1, \dots, H$, any candidate for a local solution is a stationary point or the null local solution. Therefore, if the minimum free energy contribution over all stationary points is negative, i.e., $F_h < 0$, the corresponding

stationary point is the global minimizer. With this fact, Corollary 10.6 is a straightforward deduction from Theorem 10.5. \square

Taking account of the trivial relations $c_{b_h}^2 = \widehat{\sigma}_{B_{m,h}}^2$ for $m > J$, the stationary conditions consisting of Eqs. (10.45) through (10.48), (10.52), and (10.53) for each h can be seen as a *polynomial system*, a set of polynomial equations, with $5 + J$ unknown variables, $(\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \{\widehat{\sigma}_{B_{m,h}}^2\}_{m=1}^J, c_{a_h}^2, c_{b_h}^2)$. Thus, Theorem 10.5 has decomposed the original problem with $O(M^2H^2)$ unknown variables, for which the stationary conditions are given by Eqs. (10.13) through (10.16), (10.19), and (10.20), into H subproblems with $O(J)$ unknown variables each.

Fortunately, there is a reliable numerical method to solve a polynomial system, called the *homotopy method* or *continuation method* (Drexler, 1978; Garcia and Zangwill, 1979; Gunji et al., 2004; Lee et al., 2008). It provides all isolated solutions to a system of n polynomials $f(\mathbf{x}) \equiv (f_1(\mathbf{x}), \dots, f_n(\mathbf{x})) = \mathbf{0}$ by defining a smooth set of homotopy systems with a parameter $t \in [0, 1]$, i.e., $\mathbf{g}(\mathbf{x}, t) \equiv (g_1(\mathbf{x}, t), g_2(\mathbf{x}, t), \dots, g_n(\mathbf{x}, t)) = \mathbf{0}$ such that one can continuously trace the solution path from the easiest ($t = 0$) to the target ($t = 1$). For empirical evaluation, which will be given in Section 10.8, we use HOM4PS-2.0 (Lee et al., 2008), one of the most successful polynomial system solvers.

With the homotopy method in hand, Corollary 10.6 allows us to solve the EVB learning problem (10.17) in the following way, which we call the *exact global VB solver (EGVBS)*. For each $h = 1, \dots, H$, we first find all stationary points that satisfy the polynomial system, Eqs. (10.45) through (10.48), (10.52), and (10.53). After that, we discard the prohibitive solutions with complex numbers or negative variances, and then select the stationary point giving the minimum free energy contribution F_h , defined by Eq. (10.50). The global solution is the selected stationary point if it satisfies $F_h < 0$; otherwise, the null local solution (10.54) is the global solution. Algorithm 20 summarizes the procedure of EGVBS. When the noise variance σ^2 is unknown, we conduct a naive one-dimensional search to minimize the total free energy (10.49), with EGVBS applied for every candidate value of σ^2 .

It is straightforward to modify Algorithm 20 to solve the VB learning problem (10.11), where the prior covariances $\mathbf{C}_A, \mathbf{C}_B$ are given. In this case, we should solve the polynomial system (10.45) through (10.48) in Step 3, and skip Step 6 since all local solutions are stationary points.

10.6 Approximate Global VB Solver (AGVBS)

Theorems 10.4 and 10.5 significantly reduced the complexity of the optimization problem. However, EGVBS is still not applicable to data with typical

Algorithm 20 Exact global VB solver (EGVBS) for LRSC.

-
- 1: Compute the SVD of $V = \boldsymbol{\Omega}_V^{\text{left}} \boldsymbol{\Gamma}_V \boldsymbol{\Omega}_V^{\text{right}\top}$.
 - 2: **for** $h = 1$ to H **do**
 - 3: Find all solutions of the polynomial system, Eqs. (10.45) through (10.48), (10.52), and (10.53) by the homotopy method.
 - 4: Discard prohibitive solutions with complex numbers or negative variances.
 - 5: Select the stationary point giving the smallest F_h (defined by Eq. (10.50)).
 - 6: The global solution for the h th component is the selected stationary point if it satisfies $F_h < 0$; otherwise, the null local solution (10.54) is the global solution.
 - 7: **end for**
 - 8: Compute $\widehat{U} = \boldsymbol{\Omega}_V^{\text{right}} \widehat{\mathbf{B}} \widehat{\mathbf{A}}^\top \boldsymbol{\Omega}_V^{\text{right}\top}$.
 - 9: Apply spectral clustering with the affinity matrix equal to $\text{abs}(\widehat{U}) + \text{abs}(\widehat{U}^\top)$.
-

problem sizes. This is because the homotopy method is not guaranteed to find all solutions in polynomial time in J , when the polynomial system involves $O(J)$ unknown variables.

The following simple trick further reduces the complexity and leads to an efficient approximate solver. Let us impose an additional constraint that $\gamma_m^2 \widehat{\sigma}_{B_{m,h}}^2$ are constant over $m = 1, \dots, J$, i.e.,

$$\gamma_m^2 \widehat{\sigma}_{B_{m,h}}^2 = \bar{\sigma}_{b_h}^2 \quad \text{for } m = 1, \dots, J. \quad (10.59)$$

Under this constraint, the stationary conditions for the six unknowns $(\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2, c_{a_h}^2, c_{b_h}^2)$ (for each h) become similar to the stationary conditions for fully observed MF, which allows us to obtain the following theorem:

Theorem 10.7 *Under the constraint (10.59), any stationary point of the free energy (10.50) for each h satisfies the following polynomial equation for a single variable $\widehat{\bar{\gamma}}_h \in \mathbb{R}$:*

$$\xi_6 \widehat{\bar{\gamma}}_h^6 + \xi_5 \widehat{\bar{\gamma}}_h^5 + \xi_4 \widehat{\bar{\gamma}}_h^4 + \xi_3 \widehat{\bar{\gamma}}_h^3 + \xi_2 \widehat{\bar{\gamma}}_h^2 + \xi_1 \widehat{\bar{\gamma}}_h + \xi_0 = 0, \quad (10.60)$$

where

$$\xi_6 = \frac{\phi_h^2}{\gamma_h^2}, \quad (10.61)$$

$$\xi_5 = -2 \frac{\phi_h^2 M \sigma^2}{\gamma_h^3} + \frac{2\phi_h}{\gamma_h}, \quad (10.62)$$

$$\xi_4 = \frac{\phi_h^2 M^2 \sigma^4}{\gamma_h^4} - \frac{2\phi_h(2M-J)\sigma^2}{\gamma_h^2} + 1 + \frac{\phi_h^2(M\sigma^2 - \gamma_h^2)}{\gamma_h^2}, \quad (10.63)$$

$$\xi_3 = \frac{2\phi_h M(M-J)\sigma^4}{\gamma_h^3} - \frac{2(M-J)\sigma^2}{\gamma_h} + \frac{\phi_h((M+J)\sigma^2 - \gamma_h^2)}{\gamma_h} - \frac{\phi_h^2 M\sigma^2(M\sigma^2 - \gamma_h^2)}{\gamma_h^3} + \frac{\phi_h(M\sigma^2 - \gamma_h^2)}{\gamma_h}, \quad (10.64)$$

$$\xi_2 = \frac{(M-J)^2\sigma^4}{\gamma_h^2} - \frac{\phi_h M\sigma^2((M+J)\sigma^2 - \gamma_h^2)}{\gamma_h^2} + ((M+J)\sigma^2 - \gamma_h^2) - \frac{\phi_h(M-J)\sigma^2(M\sigma^2 - \gamma_h^2)}{\gamma_h^2}, \quad (10.65)$$

$$\xi_1 = -\frac{(M-J)\sigma^2((M+J)\sigma^2 - \gamma_h^2)}{\gamma_h} + \frac{\phi_h MJ\sigma^4}{\gamma_h}, \quad (10.66)$$

$$\xi_0 = MJ\sigma^4. \quad (10.67)$$

Here $\phi_h = 1 - \frac{\gamma_h^2}{\gamma^2}$ for $\gamma^2 = (\sum_{m=1}^J \gamma_m^{-2}/J)^{-1}$. For each real solution $\widehat{\gamma}_h$ such that

$$\widehat{\gamma}_h = \widehat{\gamma}_h + \gamma_h - \frac{M\sigma^2}{\gamma_h}, \quad (10.68)$$

$$\widehat{\kappa}_h = \gamma_h^2 - (M+J)\sigma^2 - (M\sigma^2 - \gamma_h^2)\phi_h \frac{\widehat{\gamma}_h}{\gamma_h}, \quad (10.69)$$

$$\widehat{\tau}_h = \frac{1}{2MJ} \left(\widehat{\kappa}_h + \sqrt{\widehat{\kappa}_h^2 - 4MJ\sigma^4 \left(1 + \phi_h \frac{\widehat{\gamma}_h}{\gamma_h} \right)} \right), \quad (10.70)$$

$$\widehat{\delta}_h = \frac{\sigma^2}{\sqrt{\widehat{\tau}_h}} \left(\gamma_h - \frac{M\sigma^2}{\gamma_h} - \widehat{\gamma}_h \right)^{-1}, \quad (10.71)$$

are real and positive, there exists the corresponding stationary point given by

$$\left(\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2, c_{a_h}^2, c_{b_h}^2 \right) = \left(\sqrt{\widehat{\gamma}_h \widehat{\delta}_h}, \frac{\sqrt{\widehat{\gamma}_h/\widehat{\delta}_h}}{\gamma_h}, \frac{\sigma^2 \widehat{\delta}_h}{\gamma_h}, \frac{\sigma^2}{\gamma_h \widehat{\delta}_h - \phi_h \frac{\sigma^2}{\sqrt{\widehat{\tau}_h}}}, \sqrt{\widehat{\tau}_h}, \frac{\sqrt{\widehat{\tau}_h}}{\gamma_h^2} \right). \quad (10.72)$$

Given the noise variance σ^2 , computing the coefficients (10.61) through (10.67) is straightforward. Theorem 10.7 implies that the following algorithm, which we call the AGVBS, provides the global solution of the EVB learning problem (10.17) under the additional constraint (10.59). After computing the SVD of the observed matrix V , AGVBS first finds all real solutions of the sixth-order polynomial equation (10.60) by using, e.g., the “roots” command in MATLAB®, for each h . Then, it discards the prohibitive solutions such that any of Eqs. (10.68) through (10.71) gives a complex or negative number. For each of the retained solutions, AGVBS computes the corresponding stationary point by Eq. (10.72), along with the free energy contribution F_h by Eq. (10.50). Here, Eq. (10.59) is used for retrieving the original posterior variances $\{\widehat{\sigma}_{B_{m,h}}^2\}_{m=1}^J$ for B . Finally, AGVBS selects the stationary point giving the minimum free energy contribution \underline{F}_h . The global solution is the selected stationary point if it satisfies $\underline{F}_h < 0$; otherwise, the null local solution (10.54) is the global solution. Algorithm 21 summarizes the procedure of AGVBS.

Algorithm 21 Approximate global VB solver (AGVBS) for LRSC.

-
- 1: Compute the SVD of $V = \Omega_V^{\text{left}} \Gamma_V \Omega_V^{\text{right}\top}$.
 - 2: **for** $h = 1$ to H **do**
 - 3: Find all real solutions of the sixth-order polynomial equation (10.60).
 - 4: Discard prohibitive solutions such that any of Eqs. (10.68) through (10.71) gives a complex or negative number.
 - 5: Compute the corresponding stationary point by Eq. (10.72) and its free energy contribution F_h by Eq. (10.50) for each of the retained solutions.
 - 6: Select the stationary point giving the minimum free energy contribution \underline{F}_h .
 - 7: The global solution for the h th component is the selected stationary point if it satisfies $\underline{F}_h < 0$; otherwise, the null local solution (10.54) is the global solution.
 - 8: **end for**
 - 9: Compute $\widehat{U} = \Omega_V^{\text{right}} \widehat{BA}^\top \Omega_V^{\text{right}\top}$.
 - 10: Apply spectral clustering with the affinity matrix equal to $\text{abs}(\widehat{U}) + \text{abs}(\widehat{U}^\top)$.
-

As in EGVBS, a naive one-dimensional search is conducted when the noise variance σ^2 is unknown.

In Section 10.8, we show that AGVBS is practically a good alternative to the Kronecker product covariance approximation (KPCA), an approximate EVB algorithm for LRSC under the Kronecker product covariance constraint (see Section 3.4.2), in terms of accuracy and computation time.

10.7 Proof of Theorem 10.7

Let us rescale \bar{b}_h and $c_{\bar{b}_h}^2$ as follows:

$$\bar{\bar{b}}_h = \gamma_h \bar{b}_h, \quad \bar{\bar{c}}_{\bar{b}_h}^2 = \gamma_h^2 c_{\bar{b}_h}^2. \quad (10.73)$$

By substituting Eqs. (10.59) and (10.73) into Eq. (10.50), we have

$$\begin{aligned} 2F_h &= M \log \frac{c_{a_h}^2}{\bar{\sigma}_{a_h}^2} + J \log \frac{\bar{c}_{b_h}^2}{\bar{\sigma}_{b_h}^2} + \frac{1}{c_{a_h}^2} (\bar{a}_h^2 + M \bar{\sigma}_{a_h}^2) + \frac{1}{\bar{c}_{b_h}^2} \left(\bar{\bar{b}}_h^2 + J \frac{\gamma_h^2}{\underline{\gamma}^2} \bar{\sigma}_{b_h}^2 \right) \\ &\quad + \frac{1}{\sigma^2} \left(-2\gamma_h \bar{a}_h \bar{\bar{b}}_h + (\bar{a}_h^2 + M \bar{\sigma}_{a_h}^2) (\bar{\bar{b}}_h^2 + J \bar{\sigma}_{b_h}^2) \right) - (M + J) + \sum_{m=1}^J \log \frac{\gamma_m^2}{\gamma_h^2}, \end{aligned} \quad (10.74)$$

where

$$\underline{\gamma}^2 = \left(\sum_{m=1}^J \gamma_m^{-2} / J \right)^{-1}.$$

Ignoring the last two constant terms, we find that Eq. (10.74) is in almost the same form as the free energy of fully observed MF for a $J \times M$ observed matrix (see Eq. (6.43)). Only the difference is in the fourth term: $J\bar{\sigma}_{b_h}^2$ is multiplied by $\frac{\gamma_h^2}{\underline{\gamma}^2}$. Note that, as in MF, the free energy (10.74) is invariant under the following transformation:

$$\left\{ (\widehat{a}_h, \bar{\widehat{b}}_h, \widehat{\sigma}_{a_h}^2, \bar{\widehat{\sigma}}_{a_h}^2, c_{a_h}^2, \bar{c}_{b_h}^2) \right\} \rightarrow \left\{ (s_h \widehat{a}_h, s_h^{-1} \bar{\widehat{b}}_h, s_h^2 \widehat{\sigma}_{a_h}^2, s_h^{-2} \bar{\widehat{\sigma}}_{b_h}^2, s_h^2 c_{a_h}^2, s_h^{-2} \bar{c}_{b_h}^2) \right\}$$

for any $\{s_h \neq 0; h = 1, \dots, H\}$. Accordingly, we fix the ratio between c_{a_h} and \bar{c}_{b_h} to $c_{a_h}/\bar{c}_{b_h} = 1$ without loss of generality.

By differentiating the free energy (10.74) with respect to $\widehat{a}_h, \widehat{\sigma}_{a_h}^2, \bar{\widehat{b}}_h, \bar{\widehat{\sigma}}_{b_h}^2, c_{a_h}^2$, and $\bar{c}_{b_h}^2$, respectively, we obtain the following stationary conditions:

$$\widehat{a}_h = \frac{1}{\sigma^2} \gamma_h \bar{\widehat{b}}_h \widehat{\sigma}_{a_h}^2, \quad (10.75)$$

$$\widehat{\sigma}_{a_h}^2 = \sigma^2 \left(\bar{\widehat{b}}_h + J\bar{\widehat{\sigma}}_{b_h}^2 + \frac{\sigma^2}{c_{a_h}^2} \right)^{-1}, \quad (10.76)$$

$$\bar{\widehat{b}}_h = \gamma_h \widehat{a}_h \left(\widehat{\sigma}_{a_h}^2 + M\bar{\widehat{\sigma}}_{a_h}^2 + \frac{\sigma^2}{\bar{c}_{b_h}^2} \right)^{-1}, \quad (10.77)$$

$$\bar{\widehat{\sigma}}_{b_h}^2 = \sigma^2 \left(\widehat{\sigma}_{a_h}^2 + M\bar{\widehat{\sigma}}_{a_h}^2 + \frac{\sigma^2 \gamma_h^2}{\bar{c}_{b_h}^2 \underline{\gamma}^2} \right)^{-1}, \quad (10.78)$$

$$c_{a_h}^2 = \widehat{\sigma}_{a_h}^2 / M + \bar{\widehat{\sigma}}_{a_h}^2, \quad (10.79)$$

$$\bar{c}_{b_h}^2 = \bar{\widehat{b}}_h^2 / J + \frac{\gamma_h^2}{\underline{\gamma}^2} \bar{\widehat{\sigma}}_{b_h}^2. \quad (10.80)$$

Note that, unlike the case of fully observed MF, \mathbf{A} and \mathbf{B} are not symmetric, which makes analysis more involved. Apparently, if $\widehat{a}_h = 0$ or $\bar{\widehat{b}}_h = 0$, the null solution (10.54) gives the minimum $F_h \rightarrow +0$ of the free energy (10.74). In the following, we identify the positive stationary points such that $\widehat{a}_h, \bar{\widehat{b}}_h > 0$. To this end, we derive a polynomial equation with a single unknown variable from the stationary conditions (10.75) through (10.80). Let

$$\widehat{\gamma}_h = \widehat{a}_h \bar{\widehat{b}}_h, \quad (10.81)$$

$$\widehat{\delta}_h = \widehat{a}_h / \bar{\widehat{b}}_h. \quad (10.82)$$

From Eqs. (10.75) through (10.78), we obtain

$$\gamma_h^2 = \left(\bar{a}_h^2 + M\bar{\sigma}_{a_h}^2 + \frac{\sigma^2}{\bar{c}_{b_h}^2} \right) \left(\bar{b}_h^2 + J\bar{\sigma}_{b_h}^2 + \frac{\sigma^2}{\bar{c}_{a_h}^2} \right), \quad (10.83)$$

$$\gamma_h \widehat{\delta}_h^{-1} = \left(\bar{b}_h^2 + J\bar{\sigma}_{b_h}^2 + \frac{\sigma^2}{\bar{c}_{a_h}^2} \right), \quad (10.84)$$

$$\gamma_h \widehat{\delta}_h = \left(\bar{a}_h^2 + M\bar{\sigma}_{a_h}^2 + \frac{\sigma^2}{\bar{c}_{b_h}^2} \right). \quad (10.85)$$

Substituting Eq. (10.84) into Eq. (10.76) gives

$$\bar{\sigma}_{a_h}^2 = \frac{\sigma^2 \widehat{\delta}_h}{\gamma_h}. \quad (10.86)$$

Substituting Eq. (10.85) into Eq. (10.78) gives

$$\bar{\sigma}_{b_h}^2 = \frac{\sigma^2}{\gamma_h \widehat{\delta}_h - \phi_h \frac{\sigma^2}{\bar{c}_{b_h}^2}}, \quad (10.87)$$

where

$$\phi_h = 1 - \frac{\gamma_h^2}{\underline{\gamma}^2}.$$

Thus, the variances $\bar{\sigma}_{a_h}^2$ and $\bar{\sigma}_{b_h}^2$ have been written as functions of $\widehat{\delta}_h$ and $\bar{c}_{b_h}^2$.

Substituting Eqs. (10.86) and (10.87) into Eq. (10.78) gives

$$\frac{\sigma^2}{\gamma_h \widehat{\delta}_h - \phi_h \frac{\sigma^2}{\bar{c}_{b_h}^2}} \left(\bar{a}_h^2 + M \frac{\sigma^2 \widehat{\delta}_h}{\gamma_h} + \frac{\sigma^2 \gamma_h^2}{\bar{c}_{b_h}^2 \underline{\gamma}^2} \right) = \sigma^2,$$

and therefore

$$\widehat{\gamma}_h + \frac{M\sigma^2}{\gamma_h} - \gamma_h + \frac{\sigma^2}{\bar{c}_{b_h}^2} \widehat{\delta}_h^{-1} = 0.$$

Solving the preceding equation with respect to $\widehat{\delta}_h^{-1}$ gives

$$\widehat{\delta}_h^{-1} = \frac{\bar{c}_{b_h}^2}{\sigma^2} \left(\gamma_h - \frac{M\sigma^2}{\gamma_h} - \widehat{\gamma}_h \right). \quad (10.88)$$

Thus, we have obtained an expression of $\widehat{\delta}_h$ as a function of $\widehat{\gamma}_h$ and $\bar{c}_{b_h}^2$.

Substituting Eqs. (10.86) and (10.87) into Eq. (10.76) gives

$$\frac{\sigma^2 \widehat{\delta}_h}{\gamma_h} \left(\bar{b}_h^2 + J \frac{\sigma^2}{\gamma_h \widehat{\delta}_h - \phi_h \frac{\sigma^2}{\bar{c}_{b_h}^2}} + \frac{\sigma^2}{\bar{c}_{a_h}^2} \right) = \sigma^2.$$

Rearranging the previous equation with respect to $\widehat{\delta}_h^{-1}$ gives

$$(\gamma_h - \widehat{\gamma}_h) \frac{\phi_h \sigma^2}{\bar{c}_{b_h}^2 \gamma_h} \widehat{\delta}_h^{-2} + \left(\widehat{\gamma}_h + \frac{J \sigma^2}{\gamma_h} - \gamma_h - \frac{\phi_h \sigma^4}{c_{a_h}^2 \bar{c}_{b_h}^2 \gamma_h} \right) \widehat{\delta}_h^{-1} + \frac{\sigma^2}{c_{a_h}^2} = 0. \quad (10.89)$$

Substituting Eq. (10.88) into Eq. (10.89), we have

$$\begin{aligned} & \frac{\phi_h}{\gamma_h} (\widehat{\gamma}_h - \gamma_h) \left(\widehat{\gamma}_h - \left(\gamma_h - \frac{M \sigma^2}{\gamma_h} \right) \right)^2 - \frac{\sigma^4}{c_{a_h}^2 \bar{c}_{b_h}^2} \\ & + \left(\widehat{\gamma}_h - \left(\gamma_h - \frac{J \sigma^2}{\gamma_h} + \frac{\phi_h \sigma^4}{c_{a_h}^2 \bar{c}_{b_h}^2 \gamma_h} \right) \right) \left(\widehat{\gamma}_h - \left(\gamma_h - \frac{M \sigma^2}{\gamma_h} \right) \right) = 0. \end{aligned} \quad (10.90)$$

Thus we have derived an equation that includes only two unknown variables, $\widehat{\gamma}_h$ and $c_{a_h}^2 \bar{c}_{b_h}^2$.

Next we will obtain another equation that includes only $\widehat{\gamma}_h$ and $c_{a_h}^2 \bar{c}_{b_h}^2$. Substituting Eqs. (10.79) and (10.80) into Eq. (10.83), we have

$$\gamma_h^2 = \left(M c_{a_h}^2 + \frac{\sigma^2}{\bar{c}_{b_h}^2} \right) \left(J \bar{c}_{b_h}^2 + J \phi_h \bar{\sigma}_{b_h}^2 + \frac{\sigma^2}{c_{a_h}^2} \right). \quad (10.91)$$

Substituting Eq. (10.88) into Eq. (10.87) gives

$$\bar{\sigma}_{b_h}^2 = \frac{\bar{c}_{b_h}^2 \left(\gamma_h - \frac{M \sigma^2}{\gamma_h} - \widehat{\gamma}_h \right)}{\gamma_h - \phi_h \left(\gamma_h - \frac{M \sigma^2}{\gamma_h} - \widehat{\gamma}_h \right)}. \quad (10.92)$$

Substituting Eq. (10.92) into Eq. (10.91) gives

$$\gamma_h^2 = M J c_{a_h}^2 \bar{c}_{b_h}^2 + (M + J) \sigma^2 + \frac{\sigma^4}{c_{a_h}^2 \bar{c}_{b_h}^2} + J \phi_h \frac{\left(M c_{a_h}^2 \bar{c}_{b_h}^2 + \sigma^2 \right) \left(\gamma_h - \frac{M \sigma^2}{\gamma_h} - \widehat{\gamma}_h \right)}{\gamma_h - \phi_h \left(\gamma_h - \frac{M \sigma^2}{\gamma_h} - \widehat{\gamma}_h \right)}.$$

Rearranging the preceding equation with respect to $c_{a_h}^2 \bar{c}_{b_h}^2$, we have

$$M J c_{a_h}^4 \bar{c}_{b_h}^4 + \left((M + J) \sigma^2 - \gamma_h^2 + (M \sigma^2 - \gamma_h^2) \phi_h \frac{\widehat{\gamma}_h}{\gamma_h} \right) c_{a_h}^2 \bar{c}_{b_h}^2 + \sigma^4 \left(1 + \phi_h \frac{\widehat{\gamma}_h}{\gamma_h} \right) = 0, \quad (10.93)$$

where

$$\widehat{\gamma}_h = \widehat{\gamma}_h - \left(\gamma_h - \frac{M \sigma^2}{\gamma_h} \right). \quad (10.94)$$

The solution of Eq. (10.93) with respect to $c_{a_h}^2 \bar{c}_{b_h}^2$ is given by

$$c_{a_h}^2 \bar{c}_{b_h}^2 = \frac{\widehat{\kappa}_h + \sqrt{\widehat{\kappa}_h^2 - 4 M J \sigma^4 \left(1 + \phi_h \frac{\widehat{\gamma}_h}{\gamma_h} \right)}}{2 M J}, \quad (10.95)$$

where

$$\widehat{\kappa}_h = \gamma_h^2 - (M + J)\sigma^2 - (M\sigma^2 - \gamma_h^2) \phi_h \frac{\widehat{\bar{\gamma}}_h}{\gamma_h}. \quad (10.96)$$

By using Eq. (10.94), Eq. (10.90) can be rewritten as

$$\frac{1}{\gamma_h} \phi_h \left(\widehat{\bar{\gamma}}_h - \frac{M\sigma^2}{\gamma_h} \right) \widehat{\bar{\gamma}}_h^2 + \left(\widehat{\bar{\gamma}}_h - \frac{(M - J)\sigma^2}{\gamma_h} \right) \widehat{\bar{\gamma}}_h - \left(\frac{1}{\gamma_h} \phi_h \widehat{\bar{\gamma}}_h + 1 \right) \frac{\sigma^4}{c_{a_h}^2 \bar{c}_{b_h}^2} = 0. \quad (10.97)$$

Thus, we have obtained two equations, Eqs. (10.95) and (10.97), that relate two unknown variables, $\widehat{\bar{\gamma}}_h$ (or $\widehat{\gamma}_h$) and $c_{a_h}^2 \bar{c}_{b_h}^2$. Substituting Eq. (10.95) into Eq. (10.97) gives a polynomial equation involving only a single unknown variable $\widehat{\bar{\gamma}}_h$. With some algebra, we obtain Eq. (10.60).

Let

$$\widehat{\tau}_h = c_{a_h}^2 \bar{c}_{b_h}^2. \quad (10.98)$$

Since we fixed the arbitrary ratio to $c_{a_h}^2 / \bar{c}_{b_h}^2 = 1$, we have

$$c_{a_h}^2 = \sqrt{\widehat{\tau}_h}, \quad (10.99)$$

$$\bar{c}_{b_h}^2 = \sqrt{\widehat{\tau}_h}. \quad (10.100)$$

Some solutions of Eq. (10.60) have no corresponding points in the problem domain (10.18). Assume that a solution $\widehat{\bar{\gamma}}_h$ is associated with a point in the domain. Then $\widehat{\gamma}_h$ is given by Eq. (10.94), which is real and positive by its definition (10.81). $\widehat{\tau}_h$ is defined and given, respectively, by Eqs. (10.98) and (10.95), which is real and positive. $\widehat{\kappa}_h$, defined by Eq. (10.96), is also real and positive, since $\widehat{\tau}_h$ cannot be real and positive otherwise. $\widehat{\delta}_h$ is given by Eq. (10.88), which is real and positive by its definition (10.82). Finally, remembering the variable change (10.73), we can obtain Eq. (10.72) from Eqs. (10.81), (10.82), (10.86), (10.87), (10.99), and (10.100), which completes the proof of Theorem 10.7. \square

10.8 Empirical Evaluation

In this section, we empirically compare the global solvers, EGVBS (Algorithm 20) and AGVBS (Algorithm 21), with the standard iterative algorithm (Algorithm 4 in Section 3.4.2) and its approximation (Algorithm 5 in Section 3.4.2), which we here call the standard VB (SVB) iteration and the KPCA iteration, respectively. We assume that the prior covariances ($\mathbf{C}_A, \mathbf{C}_B$) and the noise

variance σ^2 are unknown and estimated from observation. We use the full-rank model (i.e., $H = \min(L, M)$), and expect EVB learning to automatically find the true rank without any parameter tuning.

Artificial Data Experiment

We first conducted an experiment with a small artificial data set (“artificial small”), on which the exact algorithms, i.e., EGVBS and the SVB iteration, are computationally tractable. Through this experiment, we can assess the accuracy of the efficient approximate solvers, i.e., AGVBS and the KPCA iteration. We randomly created $M = 75$ samples in the $L = 10$ dimensional space. We assumed $K = 2$ clusters: $M^{(1)*} = 50$ samples lie in a $H^{(1)*} = 3$ -dimensional subspace, and the other $M^{(2)*} = 25$ samples lie in a $H^{(2)*} = 1$ -dimensional subspace. For each cluster k , we independently drew $M^{(k)*}$ samples from $\text{Gauss}_{H^{(k)*}}(\mathbf{0}, 10 \cdot \mathbf{I}_{H^{(k)*}})$, and projected them onto the observed L -dimensional space by $\mathbf{R}^{(k)} \in \mathbb{R}^{L \times H^{(k)*}}$, each entry of which follows $R_{l,h}^{(k)} \sim \text{Gauss}_1(0, 1)$. Thus, we obtained a noiseless matrix $\mathbf{V}^{(k)*} \in \mathbb{R}^{L \times M^{(k)*}}$ for the k th cluster. Concatenating all clusters, $\mathbf{V}^* = (\mathbf{V}^{(1)*}, \dots, \mathbf{V}^{(K)*})$, and adding random noise subject to $\text{Gauss}_1(0, 1)$ to each entry gave an artificial observed matrix $\mathbf{V} \in \mathbb{R}^{L \times M}$, where $M = \sum_{k=1}^K M^{(k)*} = 75$. The *true* rank of \mathbf{V}^* is given by $H^* = \min(\sum_{k=1}^K H^{(k)*}, L, M) = 4$. Note that H^* is different from the rank of the observed matrix \mathbf{V} , which is almost surely equal to $J = \min(L, M) (= 10)$ under the Gaussian noise.

Figure 10.1 shows the free energy, the computation time, and the estimated rank of $\widehat{\mathbf{U}} = \widehat{\mathbf{B}}' \widehat{\mathbf{A}}'^\top$ over iterations. For the iterative methods, we show the results of 10 trials starting from different random initializations. We can see that AGVBS gives almost the same free energy as the exact methods (EGVBS and the SVB iteration). The exact methods require large computation costs: EGVBS took 621 sec to obtain the global solution, and the SVB iteration took ~ 100 sec to achieve almost the same free energy. On the other hand, the approximate methods are much faster: AGVBS took less than 1 sec, and the KPCA iteration took ~ 10 sec. Since the KPCA iteration had not converged after 250 iterations, we continued its computation until 2,500 iterations, and found that it sometimes converges to a local solution with a significantly higher free energy than the other methods. EGVBS, AGVBS, and the SVB iteration successfully found the *true* rank $H^* = 4$, while the KPCA iteration sometimes failed to find it. This difference is actually reflected to the clustering error, i.e., the misclassification rate with all possible cluster correspondences taken into account, after spectral clustering (Shi and Malik, 2000) is performed: 1.3% for EGVBS, AGVBS, and the SVB iteration, and 2.4% for the KPCA iteration.

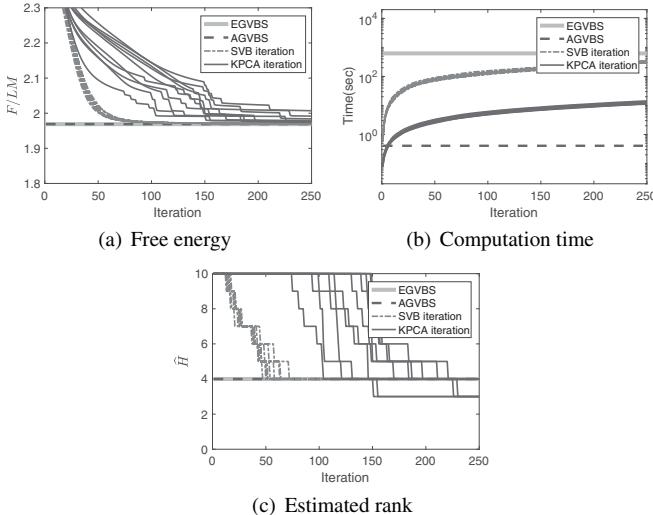


Figure 10.1 Results on the “artificial small” data set ($L = 10, M = 75, H^* = 4$). The clustering errors were 1.3% for EGVBS, AGVBS, and the SVB iteration, and 2.4% for the KPCA iteration.

Next we conducted the same experiment with a larger artificial data set (“artificial large”) ($L = 50, K = 4, (M^{(1)*}, \dots, M^{(K)*}) = (100, 50, 50, 25)$, $(H^{(1)*}, \dots, H^{(K)*}) = (2, 1, 1, 1)$), on which EGVBS and the SVB iteration are computationally intractable. Figure 10.2 shows the results with AGVBS and the KPCA iteration. The advantage in computation time is clear: AGVBS only took ~ 0.1 sec, while the KPCA iteration took more than 100 sec. The clustering errors were 4.0% for AGVBS and 11.2% for the KPCA iteration.

Benchmark Data Experiment

Finally, we applied AGVBS and the KPCA iteration to the *Hopkins 155 motion* database (Tron and Vidal, 2007). In this data set, each sample corresponds to the trajectory of a point in a video, and clustering the trajectories amounts to finding a set of rigid bodies. Figure 10.3 shows the results on the “1R2RC” ($L = 59, M = 459$) sequence.¹ We see that AGVBS gave a lower free energy with much less computation time than the KPCA iteration. Figure 10.4 shows the clustering errors on the first 20 sequences, which implies that AGVBS generally outperforms the KPCA iteration. Figure 10.4 also shows the results

¹ Peaks in the free energy curves are due to pruning. As noted in Section 3.1.1, the free energy can increase right after pruning happens, but immediately gets lower than the free energy before pruning.

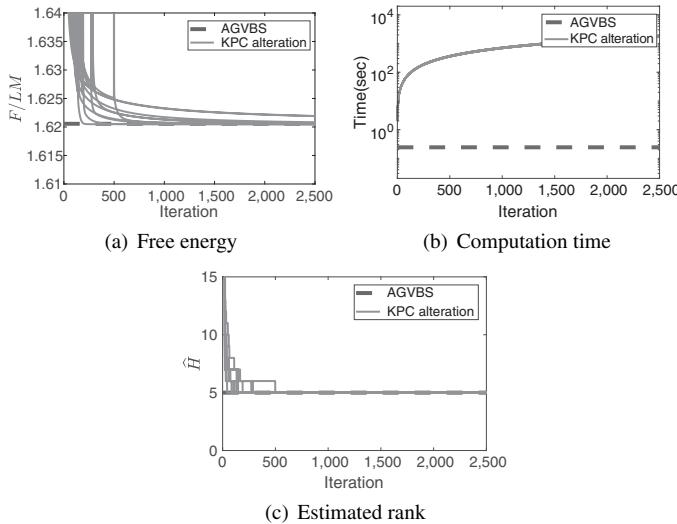


Figure 10.2 Results on the “artificial large” data set ($L = 50, M = 225, H^* = 5$). The clustering errors were 4.0% for AGVBS and 11.2% for the KPCA iteration.

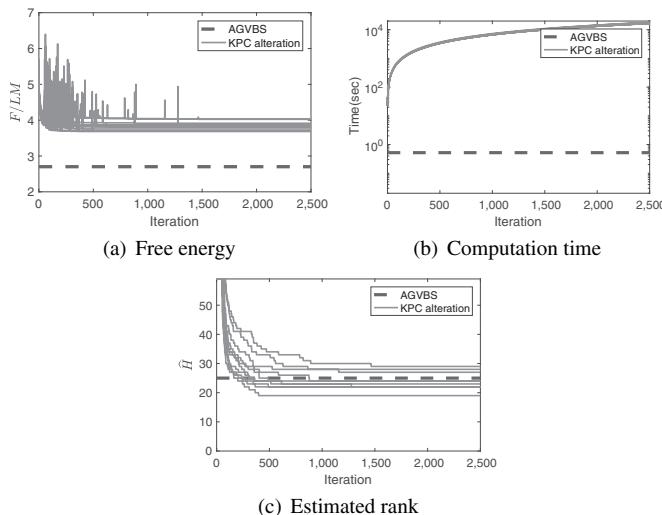


Figure 10.3 Results on the “1R2RC” sequence ($L = 59, M = 459$) of the Hopkins 155 motion database. Peaks in the free energy curves are due to pruning. The clustering errors are shown in Figure 10.4.

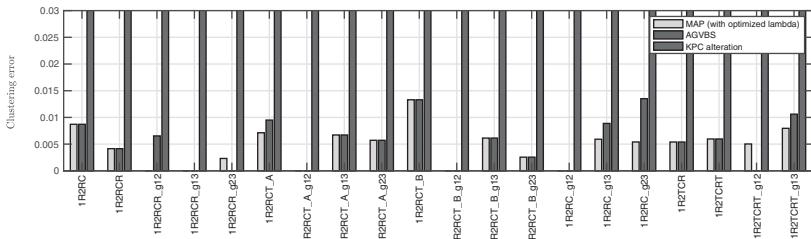


Figure 10.4 Clustering errors on the first 20 sequences of the Hopkins 155 data set.

by MAP learning (Eq. (3.87) in Section 3.4) with the tuning parameter λ optimized over the 20 sequences (i.e., we performed MAP learning with different values for λ , and selected the one giving the lowest average clustering error). We see that AGVBS performs comparably to MAP learning with optimized λ , which implies that EVB learning estimates the hyperparameters and the noise variance reasonably well.

11

Efficient Solver for Sparse Additive Matrix Factorization

In this chapter, we introduce an efficient variational Bayesian (VB) solver (Nakajima et al., 2013b) for sparse additive matrix factorization (SAMF), where the global VB solver, derived in Chapter 9, for fully observed MF is used as a subroutine.

11.1 Problem Description

The SAMF model, introduced in Section 3.5, is defined as

$$p(\mathbf{V}|\boldsymbol{\Theta}) \propto \exp\left(-\frac{1}{2\sigma^2} \left\| \mathbf{V} - \sum_{s=1}^S \mathbf{U}^{(s)} \right\|_{\text{Fro}}^2\right), \quad (11.1)$$

$$p(\{\boldsymbol{\Theta}_A^{(s)}\}_{s=1}^S) \propto \exp\left(-\frac{1}{2} \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr}\left(\mathbf{A}^{(k,s)} \mathbf{C}_A^{(k,s)-1} \mathbf{A}^{(k,s)\top}\right)\right), \quad (11.2)$$

$$p(\{\boldsymbol{\Theta}_B^{(s)}\}_{s=1}^S) \propto \exp\left(-\frac{1}{2} \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr}\left(\mathbf{B}^{(k,s)} \mathbf{C}_B^{(k,s)-1} \mathbf{B}^{(k,s)\top}\right)\right), \quad (11.3)$$

where

$$\mathbf{U}^{(s)} = \mathbf{G}(\{\mathbf{B}^{(k,s)} \mathbf{A}^{(k,s)\top}\}_{k=1}^{K^{(s)}}; \mathcal{X}^{(s)}) \quad (11.4)$$

is the s th sparse matrix factorization (SMF) term. Here $\mathbf{G}(\cdot; \mathcal{X}) : \mathbb{R}^{\prod_{k=1}^K (L^{(k)} \times M'^{(k)})} \mapsto \mathbb{R}^{L \times M}$ maps the partitioned-and-rearranged (PR) matrices $\{\mathbf{U}'^{(k)}\}_{k=1}^K$ to the target matrix $\mathbf{U} \in \mathbb{R}^{L \times M}$, based on the one-to-one map $\mathcal{X} : (k, l', m') \mapsto (l, m)$ from the indices of the entries in $\{\mathbf{U}'^{(k)}\}_{k=1}^K$ to the indices of the entries in \mathbf{U} such that

$$\left(\mathbf{G}(\{\mathbf{U}'^{(k)}\}_{k=1}^K; \mathcal{X})\right)_{l,m} = U_{l,m} = U_{\mathcal{X}(k,l',m')} = U'_{l',m'}. \quad (11.5)$$

The prior covariances of $\mathbf{A}^{(k,s)}$ and $\mathbf{B}^{(k,s)}$ are assumed to be diagonal and positive-definite:

$$\begin{aligned}\mathbf{C}_A^{(k,s)} &= \text{Diag}(c_{a_1}^{(k,s)2}, \dots, c_{a_H}^{(k,s)2}), \\ \mathbf{C}_B^{(k,s)} &= \text{Diag}(c_{b_1}^{(k,s)2}, \dots, c_{b_H}^{(k,s)2}),\end{aligned}$$

and $\boldsymbol{\Theta}$ summarizes the parameters as follows:

$$\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_A^{(s)}, \boldsymbol{\Theta}_B^{(s)}\}_{s=1}^S, \text{ where } \boldsymbol{\Theta}_A^{(s)} = \{\mathbf{A}^{(k,s)}\}_{k=1}^{K^{(s)}}, \boldsymbol{\Theta}_B^{(s)} = \{\mathbf{B}^{(k,s)}\}_{k=1}^{K^{(s)}}.$$

Under the independence constraint,

$$r(\boldsymbol{\Theta}) = \prod_{s=1}^S r_A^{(s)}(\boldsymbol{\Theta}_A^{(s)}) r_B^{(s)}(\boldsymbol{\Theta}_B^{(s)}), \quad (11.6)$$

the VB posterior minimizing the free energy can be written as

$$\begin{aligned}r(\boldsymbol{\Theta}) &= \prod_{s=1}^S \prod_{k=1}^{K^{(s)}} \left(\text{MGauss}_{M'^{(k,s)}, H'^{(k,s)}}(\mathbf{A}^{(k,s)}; \widehat{\mathbf{A}}^{(k,s)}, \widehat{\Sigma}_A^{(k,s)}) \right. \\ &\quad \cdot \text{MGauss}_{L'^{(k,s)}, H'^{(k,s)}}(\mathbf{B}^{(k,s)}; \widehat{\mathbf{B}}^{(k,s)}, \widehat{\Sigma}_B^{(k,s)}) \Big) \\ &= \prod_{s=1}^S \prod_{k=1}^{K^{(s)}} \left(\prod_{m'=1}^{M'^{(k,s)}} \text{Gauss}_{H'^{(k,s)}}(\bar{\mathbf{a}}_{m'}^{(k,s)}; \widehat{\bar{\mathbf{a}}}_{m'}^{(k,s)}, \widehat{\Sigma}_A^{(k,s)}) \right. \\ &\quad \cdot \left. \prod_{l'=1}^{L'^{(k,s)}} \text{Gauss}_{H'^{(k,s)}}(\bar{\mathbf{b}}_{l'}^{(k,s)}; \widehat{\bar{\mathbf{b}}}_{l'}^{(k,s)}, \widehat{\Sigma}_B^{(k,s)}) \right). \quad (11.7)\end{aligned}$$

The free energy can be explicitly written as

$$\begin{aligned}2F &= LM \log(2\pi\sigma^2) + \frac{\|\mathbf{V}\|_{\text{Fro}}^2}{\sigma^2} \\ &\quad + \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \left(M'^{(k,s)} \log \frac{\det(\mathbf{C}_A^{(k,s)})}{\det(\widehat{\Sigma}_A^{(k,s)})} + L'^{(k,s)} \log \frac{\det(\mathbf{C}_B^{(k,s)})}{\det(\widehat{\Sigma}_B^{(k,s)})} \right) \\ &\quad + \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr} \left\{ \mathbf{C}_A^{(k,s)-1} (\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\Sigma}_A^{(k,s)}) \right. \\ &\quad \left. + \mathbf{C}_B^{(k,s)-1} (\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\Sigma}_B^{(k,s)}) \right\} \\ &\quad + \frac{1}{\sigma^2} \text{tr} \left\{ -2\mathbf{V}^\top \left(\sum_{s=1}^S \mathbf{G}(\{\widehat{\mathbf{B}}^{(k,s)} \widehat{\mathbf{A}}^{(k,s)\top}\}_{k=1}^{K^{(s)}}; \mathcal{X}^{(s)}) \right) \right\}\end{aligned}$$

$$\begin{aligned}
& + 2 \sum_{s=1}^S \sum_{s'=s+1}^S \mathbf{G}^\top (\{\widehat{\mathbf{B}}^{(k,s)} \widehat{\mathbf{A}}^{(k,s)\top}\}_{k=1}^{K^{(s)}}; \boldsymbol{\lambda}^{(s)}) \mathbf{G} (\{\widehat{\mathbf{B}}^{(k,s')} \widehat{\mathbf{A}}^{(k,s')\top}\}_{k=1}^{K^{(s')}}; \boldsymbol{\lambda}^{(s')}) \Big\} \\
& + \frac{1}{\sigma^2} \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr} \left\{ (\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\Sigma}_A^{(k,s)}) (\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\Sigma}_B^{(k,s)}) \right\} \\
& - \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} (L'^{(k,s)} + M'^{(k,s)}) H'^{(k,s)}, \tag{11.8}
\end{aligned}$$

of which the stationary conditions are given by

$$\widehat{\mathbf{A}}^{(k,s)} = \sigma^{-2} \mathbf{Z}'^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} \widehat{\Sigma}_A^{(k,s)}, \tag{11.9}$$

$$\widehat{\Sigma}_A^{(k,s)} = \sigma^2 \left(\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\Sigma}_B^{(k,s)} + \sigma^2 \mathbf{C}_A^{(k,s)-1} \right)^{-1}, \tag{11.10}$$

$$\widehat{\mathbf{B}}^{(k,s)} = \sigma^{-2} \mathbf{Z}'^{(k,s)} \widehat{\mathbf{A}}^{(k,s)} \widehat{\Sigma}_B^{(k,s)}, \tag{11.11}$$

$$\widehat{\Sigma}_B^{(k,s)} = \sigma^2 \left(\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\Sigma}_A^{(k,s)} + \sigma^2 \mathbf{C}_B^{(k,s)-1} \right)^{-1}. \tag{11.12}$$

Here $\mathbf{Z}'^{(k,s)} \in \mathbb{R}^{L'^{(k,s)} \times M'^{(k,s)}}$ is defined as

$$Z'^{(k,s)}_{l',m'} = Z^{(s)}_{\lambda^{(s)}(k,l',m')}, \quad \text{where} \quad \mathbf{Z}^{(s)} = \mathbf{V} - \sum_{s' \neq s} \widehat{\mathbf{U}}^{(s)}. \tag{11.13}$$

The stationary conditions for the hyperparameters $\{\mathbf{C}_A^{(k,s)}, \mathbf{C}_B^{(k,s)}\}_{k=1, s=1}^{K^{(s)} S}, \sigma^2$ are given as

$$c_{a_h}^{(k,s)2} = \left\| \widehat{\mathbf{a}}_h^{(k,s)} \right\|^2 / M'^{(k,s)} + (\widehat{\Sigma}_A^{(k,s)})_{hh}, \tag{11.14}$$

$$c_{b_h}^{(k,s)2} = \left\| \widehat{\mathbf{b}}_h^{(k,s)} \right\|^2 / L'^{(k,s)} + (\widehat{\Sigma}_B^{(k,s)})_{hh}, \tag{11.15}$$

$$\begin{aligned}
\sigma^2 &= \frac{1}{LM} \left\{ \|\mathbf{V}\|_{\text{Fro}}^2 - 2 \sum_{s=1}^S \text{tr} \left(\widehat{\mathbf{U}}^{(s)\top} \left(\mathbf{V} - \sum_{s'=s+1}^S \widehat{\mathbf{U}}^{(s')} \right) \right) \right. \\
&\quad \left. + \sum_{s=1}^S \sum_{k=1}^{K^{(s)}} \text{tr} \left((\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\Sigma}_A^{(k,s)}) \cdot (\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\Sigma}_B^{(k,s)}) \right) \right\}. \tag{11.16}
\end{aligned}$$

The standard VB algorithm (Algorithm 6 in Section 3.5) iteratively applies Eqs. (11.9) through (11.12) and (11.14) through (11.16) until convergence.

11.2 Efficient Algorithm for SAMF

In this section, we derive a more efficient algorithm than the standard VB algorithm. We first present a theorem that reduces a partial SAMF problem to the (fully observed) MF problem, which can be solved analytically. Then we describe the algorithm that solves the entire SAMF problem.

11.2.1 Reduction of the Partial SAMF Problem to the MF Problem

Let us denote the mean of $\mathbf{U}^{(s)}$, defined in Eq. (11.4), over the VB posterior by

$$\begin{aligned}\widehat{\mathbf{U}}^{(s)} &= \left\langle \mathbf{U}^{(s)} \right\rangle_{r_A^{(s)}(\boldsymbol{\Theta}_A^{(s)}) r_B^{(s)}(\boldsymbol{\Theta}_B^{(s)})} \\ &= \mathbf{G} \left(\left\{ \widehat{\mathbf{B}}^{(k,s)} \widehat{\mathbf{A}}^{(k,s)\top} \right\}_{k=1}^{K^{(s)}} ; \mathcal{X}^{(s)} \right).\end{aligned}\quad (11.17)$$

Then we obtain the following theorem:

Theorem 11.1 *Given $\{\widehat{\mathbf{U}}^{(s')}\}_{s' \neq s}$ and the noise variance σ^2 , the VB posterior of $(\boldsymbol{\Theta}_A^{(s)}, \boldsymbol{\Theta}_B^{(s)}) = \{\mathbf{A}^{(k,s)}, \mathbf{B}^{(k,s)}\}_{k=1}^{K^{(s)}}$ coincides with the VB posterior of the following MF model:*

$$p(\mathbf{Z}'^{(k,s)} | \mathbf{A}^{(k,s)}, \mathbf{B}^{(k,s)}) \propto \exp \left(-\frac{1}{2\sigma^2} \|\mathbf{Z}'^{(k,s)} - \mathbf{B}^{(k,s)} \mathbf{A}^{(k,s)\top}\|_{\text{Fro}}^2 \right), \quad (11.18)$$

$$p(\mathbf{A}^{(k,s)}) \propto \exp \left(-\frac{1}{2} \text{tr} \left(\mathbf{A}^{(k,s)} \mathbf{C}_A^{(k,s)-1} \mathbf{A}^{(k,s)\top} \right) \right), \quad (11.19)$$

$$p(\mathbf{B}^{(k,s)}) \propto \exp \left(-\frac{1}{2} \text{tr} \left(\mathbf{B}^{(k,s)} \mathbf{C}_B^{(k,s)-1} \mathbf{B}^{(k,s)\top} \right) \right), \quad (11.20)$$

for each $k = 1, \dots, K^{(s)}$. Here, $\mathbf{Z}'^{(k,s)} \in \mathbb{R}^{L'^{(k,s)} \times M'^{(k,s)}}$ is defined by Eq. (11.13).

Proof Given $\{\widehat{\mathbf{U}}^{(s')}\}_{s' \neq s} = \{\{\widehat{\mathbf{B}}^{(k,s')} \widehat{\mathbf{A}}^{(k,s')\top}\}_{k=1}^{K^{(s')}}\}_{s' \neq s}$ and σ^2 as fixed constants, the free energy (11.8) can be written as a function of $\{\widehat{\mathbf{A}}^{(k,s)}, \widehat{\mathbf{B}}^{(k,s)}, \widehat{\Sigma}_A^{(k,s)}, \widehat{\Sigma}_B^{(k,s)}, \mathbf{C}_A^{(k,s)}, \mathbf{C}_B^{(k,s)}\}_{k=1}^{K^{(s)}}$ as follows:

$$2F^{(s)} \left(\{\widehat{\mathbf{A}}^{(k,s)}, \widehat{\mathbf{B}}^{(k,s)}, \widehat{\Sigma}_A^{(k,s)}, \widehat{\Sigma}_B^{(k,s)}, \mathbf{C}_A^{(k,s)}, \mathbf{C}_B^{(k,s)}\}_{k=1}^{K^{(s)}} \right) = \sum_{k=1}^{K^{(s)}} 2F^{(k,s)} + \text{const.}, \quad (11.21)$$

where

$$\begin{aligned}2F^{(k,s)} &= M'^{(k,s)} \log \frac{\det(\mathbf{C}_A^{(k,s)})}{\det(\widehat{\Sigma}_A^{(k,s)})} + L'^{(k,s)} \log \frac{\det(\mathbf{C}_B^{(k,s)})}{\det(\widehat{\Sigma}_B^{(k,s)})} \\ &+ \text{tr} \left\{ \mathbf{C}_A^{(k,s)-1} (\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\Sigma}_A^{(k,s)}) + \mathbf{C}_B^{(k,s)-1} (\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\Sigma}_B^{(k,s)}) \right\}\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{\sigma^2} \text{tr} \left\{ -2 \widehat{\mathbf{A}}^{(k,s)\top} \mathbf{Z}'^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} \right. \\
& \left. + (\widehat{\mathbf{A}}^{(k,s)\top} \widehat{\mathbf{A}}^{(k,s)} + M'^{(k,s)} \widehat{\Sigma}_A^{(k,s)}) (\widehat{\mathbf{B}}^{(k,s)\top} \widehat{\mathbf{B}}^{(k,s)} + L'^{(k,s)} \widehat{\Sigma}_B^{(k,s)}) \right\}. \quad (11.22)
\end{aligned}$$

Eq. (11.22) coincides with the free energy of the fully observed matrix factorization model (11.18) through (11.20) up to a constant (see Eq. (3.23) with \mathbf{Z}' substituted for \mathbf{V}). Therefore, the VB solution is the same. \square

Eq. (11.13) relates the entries of $\mathbf{Z}^{(s)} \in \mathbb{R}^{L \times M}$ to the entries of $\{\mathbf{Z}'^{(k,s)} \in \mathbb{R}^{L'^{(k,s)} \times M'^{(k,s)}}\}_{k=1}^{K^{(s)}}$ by using the map $\mathcal{X}^{(s)} : (k, l', m') \mapsto (l, m)$ (see Eq. (11.5) and Figure 3.3).

11.2.2 Mean Update Algorithm

Theorem 11.1 states that a partial problem of SAMF—finding the posterior of $(\mathbf{A}^{(k,s)}, \mathbf{B}^{(k,s)})$ for each $k = 1, \dots, K(s)$ given $\{\widehat{\mathbf{U}}^{(s')}_{s' \neq s}\}$ and σ^2 —can be solved by the global solver for the fully observed MF model. Specifically, we use Algorithm 16, introduced in Chapter 9, for estimating each SMF term $\widehat{\mathbf{U}}^{(s)}$ in turn. We use Eq. (11.16) for updating the noise variance σ^2 . The whole procedure, called the *mean update (MU) algorithm* (Nakajima et al., 2013b), is summarized in Algorithm 22, where $\mathbf{0}_{(d_1, d_2)}$ denotes the $d_1 \times d_2$ matrix with all entries equal to zero.

The MU algorithm is similar in spirit to the *backfitting algorithm* (Hastie and Tibshirani, 1986; D’Souza et al., 2004), where each additive term is updated to fit a dummy target. In the MU algorithm, $\mathbf{Z}^{(s)}$ defined in Eq. (11.13) corresponds to the dummy target. Although the MU algorithm globally solves a partial problem in each step, its joint global optimality over the entire

Algorithm 22 Mean update (MU) algorithm for VB SAMF.

- 1: Initialize: $\widehat{\mathbf{U}}^{(s)} \leftarrow \mathbf{0}_{(L,M)}$ for $s = 1, \dots, S$, $\sigma^2 \leftarrow \|\mathbf{V}\|_{\text{Fro}}^2 / (LM)$.
 - 2: **for** $s = 1$ to S **do**
 - 3: Compute $\mathbf{Z}'^{(k,s)} \in \mathbb{R}^{L'^{(k,s)} \times M'^{(k,s)}}$ by Eq. (11.13).
 - 4: For each partition $k = 1, \dots, K^{(s)}$, compute the solution $\mathbf{U}'^{(k,s)} = \mathbf{B}^{(k,s)} \mathbf{A}^{(k,s)\top}$ for the fully observed MF by Algorithm 16 with $\mathbf{Z}'^{(k,s)}$ as the observed matrix.
 - 5: $\widehat{\mathbf{U}}^{(s)} \leftarrow \mathbf{G}(\{\widehat{\mathbf{B}}^{(k,s)} \widehat{\mathbf{A}}^{(k,s)\top}\}_{k=1}^{K^{(s)}}; \mathcal{X}^{(s)})$.
 - 6: **end for**
 - 7: Update σ^2 by Eq. (11.16).
 - 8: Repeat 2 to 7 until convergence.
-

parameter space is not guaranteed. Nevertheless, experimental results in Section 11.3 show that the MU algorithm performs well in practice.

When Algorithm 16 is applied to the dummy target matrix $\mathbf{Z}'^{(k,s)} \in \mathbb{R}^{L'^{(k,s)} \times M'^{(k,s)}}$ in Step 4, singular value decomposition is required, which dominates the computation time. However, for many practical SMF terms, including the rowwise (3.114), the columnwise (3.115), and the elementwise (3.116) terms (as well as the segmentwise term, which will be defined for a video application in Section 11.3.2), $\mathbf{Z}'^{(k,s)}$ is a vector or scalar, i.e., $L'^{(k,s)} = 1$ or $M'^{(k,s)} = 1$ holds. In such cases, the singular value and the singular vectors are given simply by

$$\begin{aligned}\gamma_1^{(k,s)} &= \|\mathbf{Z}'^{(k,s)}\|, & \omega_{a_1}^{(k,s)} &= \mathbf{Z}'^{(k,s)} / \|\mathbf{Z}'^{(k,s)}\|, & \omega_{b_1}^{(k,s)} &= 1 & \text{if } L'^{(k,s)} = 1, \\ \gamma_1^{(k,s)} &= \|\mathbf{Z}'^{(k,s)}\|, & \omega_{a_1}^{(k,s)} &= 1, & \omega_{b_1}^{(k,s)} &= \mathbf{Z}'^{(k,s)} / \|\mathbf{Z}'^{(k,s)}\| & \text{if } M'^{(k,s)} = 1.\end{aligned}$$

11.3 Experimental Results

In this section, we experimentally show good performance of the MU algorithm (Algorithm 22) over the standard VB algorithm (Algorithm 6 in Section 3.5). We also demonstrate advantages of SAMF in its flexibility in a real-world application.

11.3.1 Mean Update vs. Standard VB

We compare the algorithms under the following model:

$$\mathbf{V} = \mathbf{U}^{\text{LRCE}} + \mathcal{E},$$

where

$$\mathbf{U}^{\text{LRCE}} = \sum_{s=1}^4 \mathbf{U}^{(s)} = \mathbf{U}^{\text{low-rank}} + \mathbf{U}^{\text{row}} + \mathbf{U}^{\text{column}} + \mathbf{U}^{\text{element}}. \quad (11.23)$$

Here, “LRCE” stands for the sum of the low-rank, rowwise, columnwise, and elementwise terms, each of which is defined in Eqs. (3.113) through (3.116). We call this model “LRCE”-SAMF. As explained in Section 3.5, “LRCE”-SAMF may be used to separate the clean signal $\mathbf{U}^{\text{low-rank}}$ from a possible rowwise sparse component (constantly broken sensors), a columnwise sparse component (accidental disturbances affecting all sensors), and an elementwise sparse component (randomly distributed spiky noise). We also evaluate “LCE”-SAMF, “LRE”-SAMF, and “LE”-SAMF, which can be regarded as generalizations of robust PCA (Candès et al., 2011; Ding et al., 2011; Babacan

et al., 2012b). Note that ‘‘LE’’-SAMF corresponds to an SAMF counterpart of robust PCA.

First, we conducted an experiment with artificial data. We assume the empirical VB scenario with unknown noise variance, i.e., all hyperparameters, $\{\mathbf{C}_A^{(k,s)}, \mathbf{C}_B^{(k,s)}\}_{k=1,s=1}^{K^{(s)} S}$, and σ^2 , are estimated from observations. We use the full-rank model ($H = \min(L, M)$) for the low-rank term $\mathbf{U}^{\text{low-rank}}$, and expect the model-induced regularization (MIR) effect (see Chapter 7) to find the true rank of $\mathbf{U}^{\text{low-rank}}$, as well as the nonzero entries in \mathbf{U}^{row} , $\mathbf{U}^{\text{column}}$, and $\mathbf{U}^{\text{element}}$.

We created an artificial data set with the data matrix size $L = 40$ and $M = 100$, and the rank $H^* = 10$ for a *true* low-rank matrix $\mathbf{U}^{\text{low-rank}*} = \mathbf{B}^* \mathbf{A}^{*\top}$. Each entry in $\mathbf{A}^* \in \mathbb{R}^{L \times H^*}$ and $\mathbf{B}^* \in \mathbb{R}^{L \times H^*}$ was drawn from $\text{Gauss}_1(0, 1)$. A *true* rowwise (columnwise) part $\mathbf{U}^{\text{row}*}$ ($\mathbf{U}^{\text{column}*}$) was created by first randomly selecting ρL rows (ρM columns) for $\rho = 0.05$, and then adding a noise subject to $\text{Gauss}_M(\mathbf{0}, \zeta \mathbf{I}_M)$ ($\text{Gauss}_L(\mathbf{0}, \zeta \mathbf{I}_L)$) for $\zeta = 100$ to each of the selected rows (columns). A *true* elementwise part $\mathbf{U}^{\text{element}*}$ was similarly created by first selecting ρLM entries and then adding a noise subject to $\text{Gauss}_1(0, \zeta)$ to each of the selected entries. Finally, an observed matrix \mathbf{V} was created by adding a noise subject to $\text{Gauss}_1(0, 1)$ to each entry of the sum $\mathbf{U}^{\text{LRCE}*}$ of the aforementioned four *true* matrices.

For the standard VB algorithm, we initialized the variational parameters and the hyperparameters in the following way: the mean parameters, $\{\widehat{\mathbf{A}}^{(k,s)}, \widehat{\mathbf{B}}^{(k,s)}\}_{k=1,s=1}^{K^{(s)} S}$, were randomly created so that each entry follows $\text{Gauss}_1(0, 1)$; the covariances, $\{\widehat{\Sigma}_A^{(k,s)}, \widehat{\Sigma}_B^{(k,s)}\}_{k=1,s=1}^{K^{(s)} S}$ and $\{\mathbf{C}_A^{(k,s)}, \mathbf{C}_B^{(k,s)}\}_{k=1,s=1}^{K^{(s)} S}$, were set to be identity; and the noise variance was set to $\sigma^2 = 1$. Note that we rescaled \mathbf{V} so that $\|\mathbf{V}\|_{\text{Fro}}^2 / (LM) = 1$, before starting iteration. We ran the standard VB algorithm 10 times, starting from different initial points, and each trial is plotted by a solid line (labeled as ‘‘Standard(iniRan)’’) in Figure 11.1.

Initialization for the MU algorithm is simple: we simply set $\widehat{\mathbf{U}}^{(s)} = \mathbf{0}_{(L,M)}$ for $s = 1, \dots, S$, and $\sigma^2 = 1$. Initialization of all other variables is not needed. Furthermore, we empirically observed that the initial value for σ^2 does not affect the result much, unless it is too small. Actually, initializing σ^2 to a large value is not harmful in the MU algorithm, because it is set to an adequate value after the first iteration with the mean parameters kept $\widehat{\mathbf{U}}^{(s)} = \mathbf{0}_{(L,M)}$. The performance of the MU algorithm is plotted by the dashed line in Figure 11.1.

Figures 11.1(a) through 11.1(c) show the free energy, the computation time, and the estimated rank, respectively, over iterations, and Figure 11.1(d) shows the reconstruction errors after 250 iterations. The reconstruction errors consist of the *overall* error $\left\| \widehat{\mathbf{U}}^{\text{LRCE}} - \mathbf{U}^{\text{LRCE}*} \right\|_{\text{Fro}}^2 / (LM)$, and the four componentwise

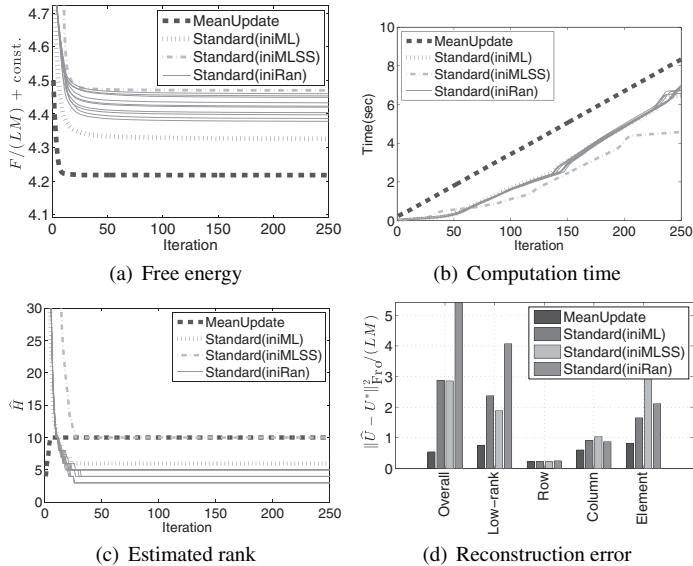


Figure 11.1 Experimental results of “LRCE”-SAMF on an artificial data set ($L = 40$, $M = 100$, $H^* = 10$, $\rho = 0.05$).

errors $\left\| \widehat{U}^{(s)} - U^{(s)*} \right\|_{\text{Fro}}^2 / (LM)$. The graphs show that the MU algorithm, whose iteration is computationally slightly more expensive than the standard VB algorithm, immediately converges to a local minimum with the free energy substantially lower than the standard VB algorithm. The estimated rank agrees with the true rank $\widehat{H} = H^* = 10$, while all 10 trials of the standard VB algorithm failed to estimate the true rank. It is also observed that the MU algorithm well reconstructs each of the four terms.

We can slightly improve the performance of the standard VB algorithm by adopting different initialization schemes. The line labeled as “Standard(iniML)” in Figure 11.1 indicates the maximum likelihood (ML) initialization, i.e., $(\widehat{\mathbf{a}}_h^{(k,s)}, \widehat{\mathbf{b}}_h^{(k,s)}) = (\gamma_h^{(k,s)1/2} \omega_{a_h}^{(k,s)}, \gamma_h^{(k,s)1/2} \omega_{b_h}^{(k,s)})$. Here, $\gamma_h^{(k,s)}$ is the h th largest singular value of the (k, s) th PR matrix $V'^{(k,s)}$ of V (such that $V'^{(k,s)} = V_{X^{(s)}(k,l',m')}^{(k,s)}$), and $\omega_{a_h}^{(k,s)}$ and $\omega_{b_h}^{(k,s)}$ are the associated right and left singular vectors. Also, we empirically found that starting from small σ^2 alleviates the local minimum problem. The line labeled as “Standard(iniMLSS)” indicates the ML initialization with $\sigma^2 = 0.0001$. We can see that this scheme successfully recovered the true rank. However, it still performs substantially worse than the MU algorithm in terms of the free energy and the reconstruction error.

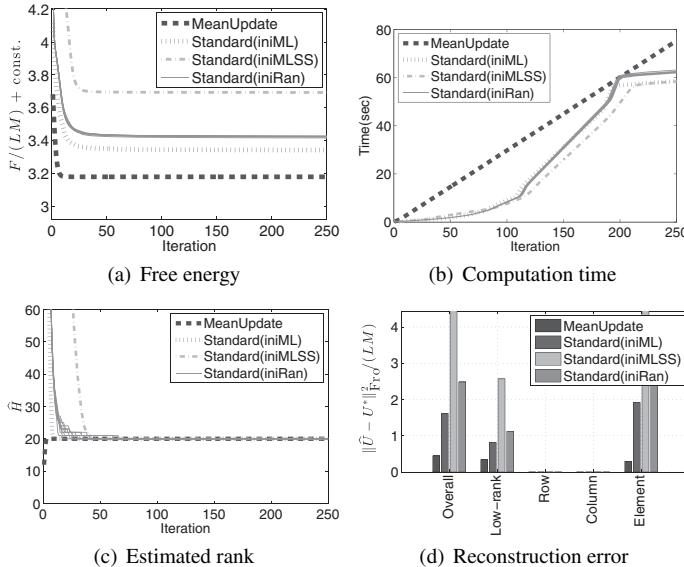


Figure 11.2 Experimental results of “LE”-SAMF on an artificial data set ($L = 100, M = 300, H^* = 20, \rho = 0.1$).

Figure 11.2 shows results of “LE”-SAMF on another artificial data set with $L = 100, M = 300, H^* = 20$, and $\rho = 0.1$. We see that the MU algorithm performs better than the standard VB algorithm. We also tested various SAMF models including “LCE”-SAMF, “LRE”-SAMF, and “LE”-SAMF under different settings for M, L, H^* , and ρ , and empirically found that the MU algorithm generally gives a better solution with lower free energy and smaller reconstruction errors than the standard VB algorithm.

Next, we conducted experiments on several data sets from the *UCI repository* (Asuncion and Newman, 2007). Since we do not know the *true* model of those data sets, we only focus on the achieved free energy. Figure 11.3 shows the free energy after convergence in “LRCE”-SAMF, “LCE”-SAMF, “LRE”-SAMF, and “LE”-SAMF. For better comparison, a constant is added so that the free energy achieved by the MU algorithm is zero. We can see a clear advantage of the MU algorithm over the standard VB algorithm.

11.3.2 Real-World Application

Finally, we demonstrate the usefulness of the flexibility of SAMF in a foreground (FG)/background (BG) video separation problem (Figure 3.5 in

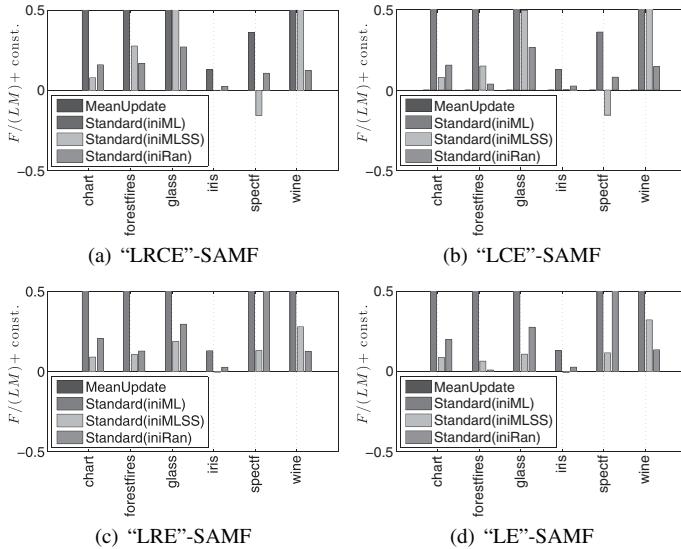


Figure 11.3 Experimental results on benchmark data sets. For better comparison, a constant is added so that the free energy achieved by the MU algorithm is zero (therefore, the bar for “MeanUpdate” is invisible).

Section 3.5). Candès et al. (2011) formed the observed matrix \mathbf{V} by stacking all pixels in each frame into each column (Figure 3.6), and applied the *robust PCA* (with “LE”-terms)—the low-rank term captures the *static* BG and the elementwise (or pixelwise) term captures the *moving* FG, e.g., people walking through. SAMF can be seen as an extension of the VB variant of robust PCA (Babacan et al., 2012b). Accordingly, we use “LE”-SAMF,

$$\mathbf{V} = \mathbf{U}^{\text{low-rank}} + \mathbf{U}^{\text{element}} + \mathcal{E},$$

as a baseline method for comparison.

The SAMF framework enables a fine-tuned design for the FG term. Assuming that pixels in an image segment with similar intensity values tend to share the same label (i.e., FG or BG), we formed a segmentwise sparse SMF term: $\mathbf{U}'^{(k)}$ for each k is a column vector consisting of all pixels in each segment. We produced an oversegmented image from each frame by using the *efficient graph-based segmentation (EGS)* algorithm (Felzenszwalb and Huttenlocher, 2004), and substituted the segmentwise sparse term for the FG term (see Figure 3.7):

$$\mathbf{V} = \mathbf{U}^{\text{low-rank}} + \mathbf{U}^{\text{segment}} + \mathcal{E}.$$

We call this model *segmentation-based SAMF* (*sSAMF*). Note that EGS is computationally very efficient: it takes less than 0.05 sec on a usual laptop to segment a 192×144 gray image. EGS has several tuning parameters, and the obtained segmentation is sensitive to some of them. However, we confirmed that *sSAMF* performs similarly with visually different segmentations obtained over a wide range of tuning parameters (see the detailed information in the section “Segmentation Algorithm”). Therefore, careful parameter tuning of EGS is not necessary for our purpose.

We compared *sSAMF* with “LE”-SAMF on the “WalkByShop1front” video from the *Caviar data set*.¹ Thanks to the Bayesian framework, all unknown parameters (except the ones for segmentation) are estimated from the data, and therefore no manual parameter tuning is required. For both models (“LE”-SAMF and *sSAMF*), we used the MU algorithm, which was shown in Section 11.3.1 to be practically more reliable than the standard VB algorithm. The original video consists of 2,360 frames, each of which is a color image with 384×288 pixels. We resized each image into 192×144 pixels, averaged over the color channels, and subsampled every 15 frames (the frame IDs are $0, 15, 30, \dots, 2355$). Thus, V is of the size of $27,684$ [pixels] \times 158 [frames]. We evaluated “LE”-SAMF and *sSAMF* on this video, and found that both models perform well (although “LE”-SAMF failed in a few frames).

In order to contrast between the two models more clearly, we created a more *difficult* video by subsampling every five frames from 1,501 to 2,000 (the frame IDs are $1501, 1506, \dots, 1996$ and V is of the size of $27,684$ [pixels] \times 100 [frames]). Since more people walked through in this period, FG/BG separation is more challenging.

Figure 11.4(a) shows one of the original frames. This is a difficult snapshot, because a person stayed at the same position for a while, which confuses separation—any object in the FG pixels is assumed to be *moving*. Figures 11.4(c) and 11.4(d) show the BG and the FG terms, respectively, obtained by “LE”-SAMF. We can see that “LE”-SAMF failed to separate the person from BG (the person is partly captured in the BG term). On the other hand, Figures 11.4(e) and 11.4(f) show the BG and the FG terms obtained by *sSAMF* based on the segmented image shown in Figure 11.4(b). We can see that *sSAMF* successfully separated the person from BG in this difficult frame. A careful look at the legs of the person reveals how segmentation helps separation—the legs form a single segment in Figure 11.4(b), and the segmentwise sparse

¹ The EC Funded CAVIAR project/IST 2001 37540, found at URL: <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>.



Figure 11.4 “LE”-SAMF vs. segmentation-based SAMF in FG/BG video separation.

term (Figure 11.4(f)) captured all pixels on the legs, while the pixelwise sparse term (Figure 11.4(d)) captured only a part of those pixels. We observed that, in all frames of the *difficult* video, as well as the *easier* one, sSAMF gave good separation, while “LE”-SAMF failed in several frames.

For reference, we applied the convex formulation of robust PCA (Candès et al., 2011), which solves the following minimization problem by the inexact augmented Lagrange multiplier (ALM) algorithm (Lin et al., 2009):

$$\min_{\boldsymbol{U}^{\text{BG}}, \boldsymbol{U}^{\text{FG}}} \|\boldsymbol{U}^{\text{BG}}\|_{\text{tr}} + \lambda \|\boldsymbol{U}^{\text{FG}}\|_1 \quad \text{s.t.} \quad \boldsymbol{V} = \boldsymbol{U}^{\text{BG}} + \boldsymbol{U}^{\text{FG}}, \quad (11.24)$$



Figure 11.5 FG/BG video separation by the convex formulation of robust PCA (11.24) for $\lambda = 0.001, 0.005, 0.025$. (top row), $\lambda = 0.001$ (middle row), and $\lambda = 0.025$ (bottom row).

where $\|\cdot\|_{\text{tr}}$ and $\|\cdot\|_1$ denote the *trace norm* and the ℓ_1 -*norm* of a matrix, respectively. Figure 11.5 shows the obtained BG and FG terms of the same frame as that in Figure 11.4 with $\lambda = 0.001, 0.005, 0.025$. We see that the performance strongly depends on the parameter value of λ , and that sSAMF gives an almost identical result (bottom row in Figure 11.4) to the best ALM result with $\lambda = 0.005$ (middle row in Figure 11.5) without any manual parameter tuning.

In the following subsections, we give detailed information on the segmentation algorithm and the computation time.

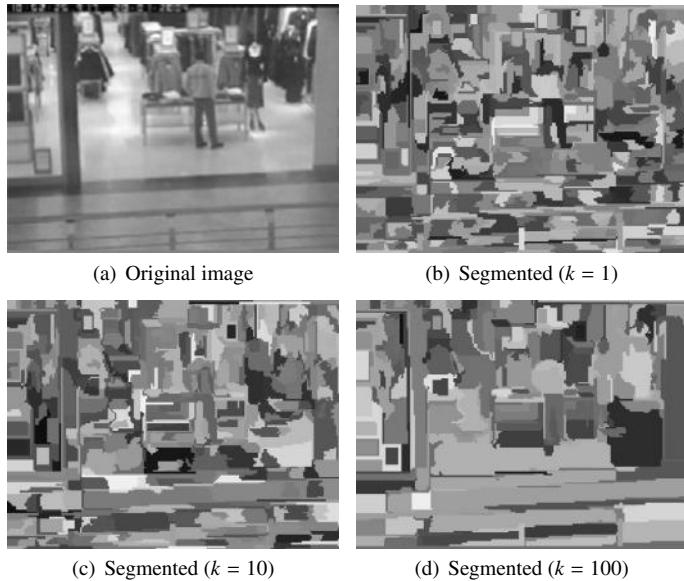


Figure 11.6 Segmented images by the efficient graph-based segmentation (EGS) algorithm with different k values. They are visually different, but with all these segmentations, sSAMF gave almost identical FB/BG separations. The original image (a) is the same frame as the one in Figure 11.4.

Segmentation Algorithm

For the EGS algorithm (Felzenszwalb and Huttenlocher, 2004), we used the code publicly available from the authors' homepage.² EGS has three tuning parameters: *sigma*, the smoothing parameter; *k*, the threshold parameter; and *minc*, the minimum segment size. Among them, *k* dominantly determines the typical size of segments (larger *k* leads to larger segments). To obtain oversegmented images for sSAMF in our experiment, we chose $k = 50$, and the other parameters are set to *sigma* = 0.5 and *minc* = 20 as recommended by the authors. We also tested other parameter setting, and observed that FG/BG separation by sSAMF performed almost equally for $1 \leq k \leq 100$, despite the visual variation of segmented images (see Figure 11.6). Overall, we empirically observed that the performance of sSAMF is not very sensitive to the selection of segmented images, unless it is highly undersegmented.

² www.cs.brown.edu/~pff/

Computation Time

The computation time for segmentation by EGS was less than 10 sec (for 100 frames). Forming the one-to-one map X took more than 80 sec (which is expected to be improved). In total, sSAMF took 600 sec on a Linux machine with Xeon X5570 (2.93GHz), while “LE”-SAMF took 700 sec. This slight reduction in computation time comes from the reduction in the number K of partitions for the FG term, and hence the number of calculations of partial analytic solutions.

12

MAP and Partially Bayesian Learning

Variational Bayesian (VB) learning generally offers a tractable approximation of Bayesian learning, and efficient iterative local search algorithms were derived for many practical models based on the conditional conjugacy (see Part II). However, in some applications, VB learning is still computationally too costly. In such cases, cruder approximation methods, where all or some of the parameters are point-estimated, with potentially less computation cost, are attractive alternatives. For example, Chu and Ghahramani (2009) applied partially Bayesian (PB) learning (introduced in Section 2.2.2), where the core tensor is integrated out and the factor matrices are point-estimated, to Tucker factorization (TF) (Carroll and Chang, 1970; Harshman, 1970; Tucker, 1996; Kolda and Bader, 2009). Mørup and Hansen (2009) applied the maximum a posteriori (MAP) learning to TF with the empirical Bayesian procedure, i.e., the hyperparameters are also estimated from observations. Their proposed empirical MAP learning, which only requires the same order of computation costs as the plain alternating least squares algorithm (Kolda and Bader, 2009), showed its model selection capability through the automatic relevance determination (ARD) property.

Motivated by the empirical success, we have analyzed PB learning and MAP learning and their empirical Bayesian variants (Nakajima et al., 2011; Nakajima and Sugiyama, 2014), which this chapter introduces. Focusing on fully observed matrix factorization (MF), we first analyze the global and local solutions of MAP learning and PB learning and their empirical Bayesian variants. This analysis theoretically reveals similarities and dissimilarities to VB learning. After that, we discuss more general cases, including MF with missing entries and TF.

12.1 Theoretical Analysis in Fully Observed MF

In this section, we formulate MAP learning and PB learning in the free energy minimization framework (Section 2.1.1) and derive analytic-form solutions.

12.1.1 Problem Description

The model likelihood and the prior of the MF model are given by

$$p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{V} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2\right), \quad (12.1)$$

$$p(\mathbf{A}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{A}\mathbf{C}_A^{-1}\mathbf{A}^\top)\right), \quad (12.2)$$

$$p(\mathbf{B}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{B}\mathbf{C}_B^{-1}\mathbf{B}^\top)\right), \quad (12.3)$$

where the prior covariance matrices are restricted to be diagonal:

$$\begin{aligned} \mathbf{C}_A &= \mathbf{Diag}(c_{a_1}^2, \dots, c_{a_H}^2), \\ \mathbf{C}_B &= \mathbf{Diag}(c_{b_1}^2, \dots, c_{b_H}^2), \end{aligned}$$

for $c_{a_h}, c_{b_h} > 0, h = 1, \dots, H$. Without loss of generality, we assume that the diagonal entries of the product $\mathbf{C}_A \mathbf{C}_B$ are arranged in nonincreasing order, i.e., $c_{a_h} c_{b_h} \geq c_{a_{h'}} c_{b_{h'}}$ for any pair $h < h'$.

As in Section 2.2.2, we treat MAP learning and PB learning as special cases of VB learning in the free energy minimization framework. The Bayes posterior is given by

$$p(\mathbf{A}, \mathbf{B}|\mathbf{V}) = \frac{p(\mathbf{V}|\mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})}{p(\mathbf{V})}, \quad (12.4)$$

which is intractable for the MF model (12.1) through (12.3). Accordingly, we approximate it by

$$\widehat{\mathbf{r}} = \underset{\mathbf{r}}{\operatorname{argmin}} F(\mathbf{r}) \quad \text{s.t.} \quad \mathbf{r}(\mathbf{A}, \mathbf{B}) \in \mathcal{G}, \quad (12.5)$$

where \mathcal{G} specifies the constraint on the approximate posterior. $F(\mathbf{r})$ is the free energy, defined as

$$\begin{aligned} F(\mathbf{r}) &= \left\langle \log \frac{r(\mathbf{A}, \mathbf{B})}{p(\mathbf{V}|\mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r(\mathbf{A}, \mathbf{B})} \\ &= \left\langle \log \frac{r(\mathbf{A}, \mathbf{B})}{p(\mathbf{A}, \mathbf{B}|\mathbf{V})} \right\rangle_{r(\mathbf{A}, \mathbf{B})} - \log p(\mathbf{V}), \end{aligned} \quad (12.6)$$

which is a monotonic function of the KL divergence $\left\langle \log \frac{r(\mathbf{A}, \mathbf{B})}{p(\mathbf{A}, \mathbf{B}|\mathbf{V})} \right\rangle_{r(\mathbf{A}, \mathbf{B})}$ to the Bayes posterior.

Constraints for MAP Learning and PB Learning

MAP learning finds the mode of the posterior distribution, which amounts to approximating the posterior with the Dirac delta function. Accordingly, solving the problem (12.5) with the following constraint gives the MAP solution:

$$r^{\text{MAP}}(\mathbf{A}, \mathbf{B}) = \delta(\mathbf{A}; \widehat{\mathbf{A}})\delta(\mathbf{B}; \widehat{\mathbf{B}}), \quad (12.7)$$

where $\delta(\boldsymbol{\mu}; \widehat{\boldsymbol{\mu}})$ denotes the (*pseudo*-)Dirac delta function located at $\widehat{\boldsymbol{\mu}}$.¹

Under the MAP constraint (12.7), the free energy (12.6) is written as

$$\begin{aligned} F^{\text{MAP}}(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) &= \left\langle \log \frac{\delta(\mathbf{A}; \widehat{\mathbf{A}})\delta(\mathbf{B}; \widehat{\mathbf{B}})}{p(\mathbf{V}|\mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})} \right\rangle_{\delta(\mathbf{A}; \widehat{\mathbf{A}})\delta(\mathbf{B}; \widehat{\mathbf{B}})} \\ &= -\log p(\mathbf{V}|\widehat{\mathbf{A}}, \widehat{\mathbf{B}})p(\widehat{\mathbf{A}})p(\widehat{\mathbf{B}}) + \chi_A + \chi_B, \end{aligned} \quad (12.8)$$

where

$$\chi_A = \left\langle \log \delta(\mathbf{A}; \widehat{\mathbf{A}}) \right\rangle_{\delta(\mathbf{A}; \widehat{\mathbf{A}})}, \quad \chi_B = \left\langle \log \delta(\mathbf{B}; \widehat{\mathbf{B}}) \right\rangle_{\delta(\mathbf{B}; \widehat{\mathbf{B}})} \quad (12.9)$$

are the negative entropies of the pseudo-Dirac delta functions.

PB learning is a strategy to *analytically* integrate out as many parameters as possible, and the rest are point-estimated. In the MF model, a natural choice is to integrate \mathbf{A} out and point-estimate \mathbf{B} , which we call PB-A learning, or to integrate \mathbf{B} out and point-estimate \mathbf{A} , which we call PB-B learning. Their solutions can be obtained by solving the problem (12.5) with the following constraints, respectively:

$$r^{\text{PB-A}}(\mathbf{A}, \mathbf{B}) = r_A^{\text{PB}}(\mathbf{A})\delta(\mathbf{B}; \widehat{\mathbf{B}}), \quad (12.10)$$

$$r^{\text{PB-B}}(\mathbf{A}, \mathbf{B}) = \delta(\mathbf{A}; \widehat{\mathbf{A}})r_B^{\text{PB}}(\mathbf{B}). \quad (12.11)$$

Under the PB-A constraint (12.10), the free energy (12.6) is written as

$$\begin{aligned} F^{\text{PB-A}}(r_A^{\text{PB}}, \widehat{\mathbf{B}}) &= \left\langle \log \frac{r_A^{\text{PB}}(\mathbf{A})}{p(\mathbf{V}|\mathbf{A}, \widehat{\mathbf{B}})p(\mathbf{A})p(\widehat{\mathbf{B}})} \right\rangle_{r_A^{\text{PB}}(\mathbf{A})} + \chi_B \\ &= \left\langle \log \frac{r_A^{\text{PB}}(\mathbf{A})}{p(\mathbf{A}|\mathbf{V}, \widehat{\mathbf{B}})} \right\rangle_{r_A^{\text{PB}}(\mathbf{A})} - \log p(\mathbf{V}|\widehat{\mathbf{B}})p(\widehat{\mathbf{B}}) + \chi_B, \end{aligned} \quad (12.12)$$

¹ By the *pseudo*-Dirac delta function, we mean an extremely localized density function, e.g., $\delta(\mathbf{A}; \widehat{\mathbf{A}}) \propto \exp(-\|\mathbf{A} - \widehat{\mathbf{A}}\|_{\text{Fro}}^2/(2\varepsilon^2))$ with a very small but strictly positive variance $\varepsilon^2 > 0$, such that its tail effect can be ignored, while its negative entropy $\chi_A = \langle \log \delta(\mathbf{A}; \widehat{\mathbf{A}}) \rangle_{\delta(\mathbf{A}; \widehat{\mathbf{A}})}$ remains finite.

where

$$p(\mathbf{A}|\mathbf{V}, \widehat{\mathbf{B}}) = \frac{p(\mathbf{V}|\mathbf{A}, \widehat{\mathbf{B}})p(\mathbf{A})}{p(\mathbf{V}|\widehat{\mathbf{B}})}, \quad (12.13)$$

$$\text{and} \quad p(\mathbf{V}|\widehat{\mathbf{B}}) = \left\langle p(\mathbf{V}|\mathbf{A}, \widehat{\mathbf{B}}) \right\rangle_{p(\mathbf{A})} \quad (12.14)$$

are the posterior distribution with respect to \mathbf{A} (given $\widehat{\mathbf{B}}$) and the marginal distribution, respectively. Note that Eq. (12.12) is a functional of r_A^{PB} and $\widehat{\mathbf{B}}$, and χ_B is a constant with respect to them.

Since only the first term depends on r_A^{PB} , on which no restriction is imposed, Eq. (12.12) is minimized when

$$r_A^{\text{PB}}(\mathbf{A}) = p(\mathbf{A}|\mathbf{V}, \widehat{\mathbf{B}}) \quad (12.15)$$

for any $\widehat{\mathbf{B}}$. With Eq. (12.15), the first term in Eq. (12.12) vanishes, and thus the estimator for $\widehat{\mathbf{B}}$ is given by

$$\widehat{\mathbf{B}}^{\text{PB-A}} = \underset{\widehat{\mathbf{B}}}{\operatorname{argmin}} \tilde{F}^{\text{PB-A}}(\widehat{\mathbf{B}}), \quad (12.16)$$

where

$$\tilde{F}^{\text{PB-A}}(\widehat{\mathbf{B}}) \equiv \min_{r_A^{\text{PB}}} F^{\text{PB-A}}(r_A^{\text{PB}}, \widehat{\mathbf{B}}) = -\log p(\mathbf{V}|\widehat{\mathbf{B}})p(\widehat{\mathbf{B}}) + \chi_B. \quad (12.17)$$

The process to compute $\tilde{F}^{\text{PB-A}}(\widehat{\mathbf{B}})$ in Eq. (12.17) corresponds to integrating \mathbf{A} out based on the conditional conjugacy. The probabilistic PCA, introduced in Section 3.1.2, was originally proposed with PB-A learning (Tipping and Bishop, 1999).

In the same way, we can obtain the approximate posterior under the PB-B constraint (12.11) as follows:

$$r_B^{\text{PB}}(\mathbf{B}) = p(\mathbf{B}|\mathbf{V}, \widehat{\mathbf{A}}), \quad (12.18)$$

$$\widehat{\mathbf{A}}^{\text{PB-B}} = \underset{\widehat{\mathbf{A}}}{\operatorname{argmin}} \tilde{F}^{\text{PB-B}}(\widehat{\mathbf{A}}), \quad (12.19)$$

where

$$p(\mathbf{B}|\mathbf{V}, \widehat{\mathbf{A}}) = \frac{p(\mathbf{V}|\widehat{\mathbf{A}}, \mathbf{B})p(\mathbf{B})}{p(\mathbf{V}|\widehat{\mathbf{A}})}, \quad (12.20)$$

$$p(\mathbf{V}|\widehat{\mathbf{A}}) = \left\langle p(\mathbf{V}|\widehat{\mathbf{A}}, \mathbf{B}) \right\rangle_{p(\mathbf{B})}, \quad (12.21)$$

$$\tilde{F}^{\text{PB-B}}(\widehat{\mathbf{A}}) \equiv \min_{r_B^{\text{PB}}} F^{\text{PB-B}}(r_B^{\text{PB}}, \widehat{\mathbf{A}}) = -\log p(\mathbf{V}|\widehat{\mathbf{A}})p(\widehat{\mathbf{A}}) + \chi_A. \quad (12.22)$$

We define PB learning as one of PB-A learning and PB-B learning giving a lower free energy. Namely,

$$r^{\text{PB}}(\mathbf{A}, \mathbf{B}) = \begin{cases} r^{\text{PB-A}}(\mathbf{A}, \mathbf{B}) & \text{if } \min F^{\text{PB-A}}(r_A^{\text{PB}}, \widehat{\mathbf{B}}) \leq \min F^{\text{PB-B}}(r_B^{\text{PB}}, \widehat{\mathbf{A}}), \\ r^{\text{PB-B}}(\mathbf{A}, \mathbf{B}) & \text{otherwise.} \end{cases}$$

Free Energies for MAP Learning and PB Learning

Apparently, the constraint (12.7) for MAP learning and the constraints (12.10) and (12.11) for PB learning forces independence between \mathbf{A} and \mathbf{B} , and therefore, they are stronger than the independence constraint

$$r^{\text{VB}}(\mathbf{A}, \mathbf{B}) = r_A^{\text{VB}}(\mathbf{A})r_B^{\text{VB}}(\mathbf{B}) \quad (12.23)$$

for VB learning. In Chapter 6, we showed that the VB posterior under the independence constraint (12.23) is in the following Gaussian form:

$$r_A(\mathbf{A}) = \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \mathbf{I}_M \otimes \widehat{\Sigma}_A) \propto \exp\left(-\frac{\text{tr}((\mathbf{A} - \widehat{\mathbf{A}})\widehat{\Sigma}_A^{-1}(\mathbf{A} - \widehat{\mathbf{A}})^\top)}{2}\right), \quad (12.24)$$

$$r_B(\mathbf{B}) = \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \mathbf{I}_L \otimes \widehat{\Sigma}_B) \propto \exp\left(-\frac{\text{tr}((\mathbf{B} - \widehat{\mathbf{B}})\widehat{\Sigma}_B^{-1}(\mathbf{B} - \widehat{\mathbf{B}})^\top)}{2}\right), \quad (12.25)$$

where the posterior covariances, $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$, are diagonal. Furthermore, Eqs. (12.24) and (12.25) can be the pseudo-Dirac delta functions by setting the posterior covariances to $\widehat{\Sigma}_A = \varepsilon^2 \mathbf{I}_H$ and $\widehat{\Sigma}_B = \varepsilon^2 \mathbf{I}_L$, respectively, for a very small $\varepsilon^2 > 0$.

Consequently, the MAP and the PB solutions can be obtained by minimizing the free energy for VB learning with posterior covariances clipped to $\varepsilon^2 \mathbf{I}_H$, according to the corresponding constraint. Namely, we start from the free energy expression (6.42) for VB learning, i.e.,

$$2F = LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^L \gamma_h^2}{\sigma^2} + \sum_{h=1}^H 2F_h, \quad (12.26)$$

where

$$\begin{aligned} 2F_h &= M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2}{c_{b_h}^2} \\ &\quad - (L + M) + \frac{-2\widehat{a}_h\widehat{b}_h\gamma_h + (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2)(\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2)}{\sigma^2}, \end{aligned} \quad (12.27)$$

and set

$$\widehat{\sigma}_{a_h}^2 = \varepsilon^2 \quad h = 1, \dots, H, \quad (12.28)$$

for MAP learning and PB-B learning, and

$$\widehat{\sigma}_{b_h}^2 = \varepsilon^2 \quad h = 1, \dots, H, \quad (12.29)$$

for MAP learning and PB-A learning. Here,

$$\begin{aligned}\widehat{\mathbf{A}} &= (\widehat{a}_1 \omega_{a_1}, \dots, \widehat{a}_H \omega_{a_H}), \\ \widehat{\mathbf{B}} &= (\widehat{b}_1 \omega_{b_1}, \dots, \widehat{b}_H \omega_{b_H}), \\ \widehat{\Sigma}_A &= \text{Diag}(\widehat{\sigma}_{a_1}^2, \dots, \widehat{\sigma}_{a_H}^2), \\ \widehat{\Sigma}_B &= \text{Diag}(\widehat{\sigma}_{b_1}^2, \dots, \widehat{\sigma}_{b_H}^2),\end{aligned}$$

and

$$\mathbf{V} = \sum_{h=1}^L \gamma_h \omega_{b_h} \omega_{a_h}^\top$$

is the singular value decomposition (SVD) of \mathbf{V} .

Thus, the free energies for MAP learning, PB-A learning, and PB-B learning are given by Eq. (12.26) for

$$\begin{aligned}2F_h^{\text{MAP}} &= M \log c_{a_h}^2 + L \log c_{b_h}^2 + \frac{\widehat{a}_h^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{c_{b_h}^2} + \frac{-2\widehat{a}_h \widehat{b}_h \gamma_h + \widehat{a}_h^2 \widehat{b}_h^2}{\sigma^2} \\ &\quad - (L+M) + (L+M)\chi,\end{aligned} \quad (12.30)$$

$$\begin{aligned}2F_h^{\text{PB-A}} &= M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log c_{b_h}^2 + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{c_{b_h}^2} + \frac{-2\widehat{a}_h \widehat{b}_h \gamma_h + (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2) \widehat{b}_h^2}{\sigma^2} \\ &\quad - (L+M) + L\chi,\end{aligned} \quad (12.31)$$

$$\begin{aligned}2F_h^{\text{PB-B}} &= M \log c_{a_h}^2 + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{\widehat{a}_h^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2}{c_{b_h}^2} + \frac{-2\widehat{a}_h \widehat{b}_h \gamma_h + \widehat{a}_h^2 (\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2)}{\sigma^2} \\ &\quad - (L+M) + M\chi,\end{aligned} \quad (12.32)$$

respectively, where

$$\chi = -\log \varepsilon^2 \quad (12.33)$$

is a large positive constant corresponding to the negative entropy of the one-dimensional pseudo-Dirac delta function.

As in VB learning, the free energy (12.26) is separable for each singular component as long as the noise variance σ^2 is treated as a constant. Therefore, the variational parameters $(\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2)$ and the prior covariances $(c_{a_h}^2, c_{b_h}^2)$ for the h th component can be estimated by minimizing F_h .

12.1.2 Global Solutions

In this section, we derive the global minimizers of the free energies, (12.30) through (12.32), and analyze their behavior.

Global MAP and PB Solutions

By minimizing the MAP free energy (12.30), we can obtain the global solution for MAP learning, given the hyperparameters $\mathbf{C}_A, \mathbf{C}_B, \sigma^2$ treated as fixed constants. Let

$$\mathbf{U} = \mathbf{B}\mathbf{A}^\top$$

be the target low-rank matrix.

Theorem 12.1 *Given $\mathbf{C}_A, \mathbf{C}_B \in \mathbb{D}_{++}^H$, and $\sigma^2 \in \mathbb{R}_{++}$, the MAP solution of the MF model (12.1) through (12.3) is given by*

$$\widehat{\mathbf{U}}^{\text{MAP}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{MAP}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \quad \text{where} \quad \widehat{\gamma}_h^{\text{MAP}} = \begin{cases} \check{\gamma}_h^{\text{MAP}} & \text{if } \gamma_h \geq \underline{\gamma}_h^{\text{MAP}}, \\ 0 & \text{otherwise,} \end{cases} \quad (12.34)$$

where

$$\underline{\gamma}_h^{\text{MAP}} = \frac{\sigma^2}{c_{a_h} c_{b_h}}, \quad (12.35)$$

$$\check{\gamma}_h^{\text{MAP}} = \gamma_h - \frac{\sigma^2}{c_{a_h} c_{b_h}}. \quad (12.36)$$

Proof Eq. (12.30) can be written as a function of \widehat{a}_h and \widehat{b}_h as

$$\begin{aligned} 2F_h^{\text{MAP}} &= \frac{\widehat{a}_h^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{c_{b_h}^2} + \frac{-2\widehat{a}_h \widehat{b}_h \gamma_h + \widehat{a}_h^2 \widehat{b}_h^2}{\sigma^2} + \text{const.} \\ &= \left(\frac{\widehat{a}_h}{c_{a_h}} - \frac{\widehat{b}_h}{c_{b_h}} \right)^2 + \frac{\left(\widehat{a}_h \widehat{b}_h - \left(\gamma_h - \frac{\sigma^2}{c_{a_h} c_{b_h}} \right) \right)^2}{\sigma^2} + \text{const.} \end{aligned} \quad (12.37)$$

Noting that the first two terms are nonnegative, we find that, if $\gamma_h > \underline{\gamma}_h^{\text{MAP}}$, Eq. (12.37) is minimized when

$$\check{\gamma}_h^{\text{MAP}} \equiv \widehat{a}_h \widehat{b}_h = \gamma_h - \frac{\sigma^2}{c_{a_h} c_{b_h}}, \quad (12.38)$$

$$\widehat{\delta}_h^{\text{MAP}} \equiv \frac{\widehat{a}_h}{\widehat{b}_h} = \frac{c_{a_h}}{c_{b_h}}. \quad (12.39)$$

Otherwise, it is minimized when

$$\begin{aligned}\widehat{a}_h &= 0, \\ \widehat{b}_h &= 0,\end{aligned}$$

which completes the proof. \square

Eqs. (12.38) and (12.39) immediately lead to the following corollary:

Corollary 12.2 *The MAP posterior is given by*

$$r^{\text{MAP}}(\mathbf{A}, \mathbf{B}) = \prod_{h=1}^H \delta(\mathbf{a}_h; \widehat{a}_h \boldsymbol{\omega}_{a_h}) \prod_{h=1}^H \delta(\mathbf{b}_h; \widehat{b}_h \boldsymbol{\omega}_{b_h}), \quad (12.40)$$

with the following estimators: if $\gamma_h > \underline{\gamma}_h^{\text{MAP}}$,

$$\widehat{a}_h = \pm \sqrt{\check{\gamma}_h^{\text{MAP}} \widehat{\delta}_h^{\text{MAP}}}, \quad \widehat{b}_h = \pm \sqrt{\frac{\check{\gamma}_h^{\text{MAP}}}{\widehat{\delta}_h^{\text{MAP}}}}, \quad (12.41)$$

$$\text{where } \widehat{\delta}_h^{\text{MAP}} \left(\equiv \frac{\widehat{a}_h}{\widehat{b}_h} \right) = \frac{c_{a_h}}{c_{b_h}}, \quad (12.42)$$

and otherwise

$$\widehat{a}_h = 0, \quad \widehat{b}_h = 0. \quad (12.43)$$

Similarly, by minimizing the PB-A free energy (12.31) and the PB-B free energy (12.32) and comparing them, we can obtain the global solution for PB learning:

Theorem 12.3 *Given $\mathbf{C}_A, \mathbf{C}_B \in \mathbb{D}_{++}^H$, and $\sigma^2 \in \mathbb{R}_{++}$, the PB solution of the MF model (12.1) through (12.3) is given by*

$$\widehat{\mathbf{U}}^{\text{PB}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{PB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \quad \text{where} \quad \widehat{\gamma}_h^{\text{PB}} = \begin{cases} \check{\gamma}_h^{\text{PB}} & \text{if } \gamma_h \geq \underline{\gamma}_h^{\text{PB}}, \\ 0 & \text{otherwise,} \end{cases} \quad (12.44)$$

where

$$\underline{\gamma}_h^{\text{PB}} = \sigma \sqrt{\max(L, M) + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}}, \quad (12.45)$$

$$\check{\gamma}_h^{\text{PB}} = \left(1 - \frac{\sigma^2 \left(\max(L, M) + \sqrt{\max(L, M)^2 + 4 \frac{\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)}{2 \gamma_h^2} \right) \gamma_h. \quad (12.46)$$

Corollary 12.4 *The PB posterior is given by*

$$r^{\text{PB}}(\mathbf{A}, \mathbf{B}) = \begin{cases} r^{\text{PB-A}}(\mathbf{A}, \mathbf{B}) & \text{if } L < M, \\ r^{\text{PB-A}}(\mathbf{A}, \mathbf{B}) \text{ or } r^{\text{PB-B}}(\mathbf{A}, \mathbf{B}) & \text{if } L = M, \\ r^{\text{PB-B}}(\mathbf{A}, \mathbf{B}) & \text{if } L > M, \end{cases} \quad (12.47)$$

where $r^{\text{PB-A}}(\mathbf{A}, \mathbf{B})$ and $r^{\text{PB-B}}(\mathbf{A}, \mathbf{B})$ are the PB-A posterior and the PB-B posterior, respectively, given as follows. The PB-A posterior is given by

$$r^{\text{PB-A}}(\mathbf{A}, \mathbf{B}) = \prod_{h=1}^H \text{Gauss}_M(\mathbf{a}_h; \widehat{\mathbf{a}}_h \boldsymbol{\omega}_{a_h}, \widehat{\sigma}_{a_h}^2 \mathbf{I}_M) \prod_{h=1}^H \delta(\mathbf{b}_h; \widehat{\mathbf{b}}_h \boldsymbol{\omega}_{b_h}), \quad (12.48)$$

with the following estimators: if

$$\gamma_h > \underline{\gamma}_h^{\text{PB-A}} \equiv \sigma \sqrt{M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}}, \quad (12.49)$$

then

$$\widehat{\mathbf{a}}_h = \pm \sqrt{\check{\gamma}_h^{\text{PB-A}} \widehat{\delta}_h^{\text{PB-A}}}, \quad \widehat{\mathbf{b}}_h = \pm \sqrt{\frac{\check{\gamma}_h^{\text{PB-A}}}{\widehat{\delta}_h^{\text{PB-A}}}}, \quad \widehat{\sigma}_{a_h}^2 = \frac{\sigma^2}{\check{\gamma}_h^{\text{PB-A}} / \widehat{\delta}_h^{\text{PB-A}} + \sigma^2 / c_{a_h}^2}, \quad (12.50)$$

where

$$\check{\gamma}_h^{\text{PB-A}} \left(\equiv \widehat{\mathbf{a}}_h \widehat{\mathbf{b}}_h \right) = \left(1 - \frac{\sigma^2 \left(M + \sqrt{M^2 + 4 \frac{\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)}{2 \gamma_h^2} \right) \gamma_h, \quad (12.51)$$

$$\widehat{\delta}_h^{\text{PB-A}} \left(\equiv \frac{\widehat{\mathbf{a}}_h}{\widehat{\mathbf{b}}_h} \right) = \frac{c_{a_h}^2 \left(M + \sqrt{M^2 + 4 \frac{\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)}{2 \gamma_h}, \quad (12.52)$$

and otherwise

$$\widehat{\mathbf{a}}_h = 0, \quad \widehat{\mathbf{b}}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = c_{a_h}^2. \quad (12.53)$$

The PB-B posterior is given by

$$r^{\text{PB-B}}(\mathbf{A}, \mathbf{B}) = \prod_{h=1}^H \delta(\mathbf{a}_h; \widehat{\mathbf{a}}_h \boldsymbol{\omega}_{a_h}) \prod_{h=1}^H \text{Gauss}_L(\mathbf{b}_h; \widehat{\mathbf{b}}_h \boldsymbol{\omega}_{b_h}, \widehat{\sigma}_{b_h}^2 \mathbf{I}_L), \quad (12.54)$$

with the following estimators: if

$$\gamma_h > \underline{\gamma}_h^{\text{PB-B}} \equiv \sigma \sqrt{L + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}}, \quad (12.55)$$

then

$$\widehat{a}_h = \pm \sqrt{\check{\gamma}_h^{\text{PB-B}} \widehat{\delta}_h^{\text{PB-B}}}, \quad \widehat{b}_h = \pm \sqrt{\frac{\check{\gamma}_h^{\text{PB-B}}}{\widehat{\delta}_h^{\text{PB-B}}}}, \quad \widehat{\sigma}_{b_h}^2 = \frac{\sigma^2}{\check{\gamma}_h^{\text{PB-B}} \widehat{\delta}_h^{\text{PB-B}} + \sigma^2 / c_{b_h}^2}, \quad (12.56)$$

where

$$\check{\gamma}_h^{\text{PB-B}} \left(\equiv \widehat{a}_h \widehat{b}_h \right) = \left(1 - \frac{\sigma^2 \left(L + \sqrt{L^2 + 4 \frac{\check{\gamma}_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)}{2 \check{\gamma}_h^2} \right) \gamma_h, \quad (12.57)$$

$$\widehat{\delta}_h^{\text{PB-B}} \left(\equiv \frac{\widehat{a}_h}{\widehat{b}_h} \right) = \left(\frac{c_{b_h}^2 \left(L + \sqrt{L^2 + 4 \frac{\check{\gamma}_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)}{2 \check{\gamma}_h} \right)^{-1}, \quad (12.58)$$

and otherwise

$$\widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad \widehat{\sigma}_{b_h}^2 = c_{b_h}^2. \quad (12.59)$$

Note that, when $L = M$, the choice from the PB-A and the PB-B posteriors depends on the prior covariances \mathbf{C}_A and \mathbf{C}_B . However, as long as the estimator for the target low-rank matrix \mathbf{U} is concerned, the choice does not matter, as Theorem 12.3 states. This is because

$$\underline{\gamma}_h^{\text{PB-A}} = \check{\gamma}_h^{\text{PB-B}} \quad \text{and} \quad \check{\gamma}_h^{\text{PB-A}} = \check{\gamma}_h^{\text{PB-B}} \quad \text{when} \quad L = M,$$

as Corollary 12.4 implies.

Proofs of Theorem 12.3 and Corollary 12.4

We first derive the PB-A solution by minimizing the corresponding free energy (12.31):

$$2F_h^{\text{PB-A}} = M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log c_{b_h}^2 + \frac{\widehat{a}_h^2 + M \widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{c_{b_h}^2} + \frac{-2\widehat{a}_h \widehat{b}_h \gamma_h + (\widehat{a}_h^2 + M \widehat{\sigma}_{a_h}^2) \widehat{b}_h^2}{\sigma^2} - (L + M) + L\chi. \quad (12.60)$$

As a function of \widehat{a}_h (treating \widehat{b}_h and $\widehat{\sigma}_{a_h}^2$ as fixed constants), Eq. (12.60) can be written as

$$2F_h^{\text{PB-A}}(\widehat{a}_h) = \frac{\widehat{a}_h^2}{c_{a_h}^2} + \frac{-2\widehat{a}_h \widehat{b}_h \gamma_h + \widehat{a}_h^2 \widehat{b}_h^2}{\sigma^2} + \text{const.}$$

$$\begin{aligned}
&= \frac{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2}{\sigma^2} \left(\widehat{a}_h^2 - 2 \frac{\widehat{b}_h \gamma_h}{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2} \widehat{a}_h \right) + \text{const.} \\
&= \frac{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2}{\sigma^2} \left(\widehat{a}_h - \frac{\widehat{b}_h \gamma_h}{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2} \right)^2 + \text{const.,}
\end{aligned}$$

which is minimized when

$$\widehat{a}_h = \frac{\widehat{b}_h \gamma_h}{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2}. \quad (12.61)$$

As a function of $\widehat{\sigma}_{a_h}^2$ (treating \widehat{a}_h and \widehat{b}_h as fixed constants), Eq. (12.60) can be written as

$$\begin{aligned}
2F_h^{\text{PB-A}}(\widehat{\sigma}_{a_h}^2) &= M \left(-\log \widehat{\sigma}_{a_h}^2 + \left(\frac{1}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{\sigma^2} \right) \widehat{\sigma}_{a_h}^2 \right) + \text{const.} \\
&= M \left(-\log \left(\frac{1}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{\sigma^2} \right) \widehat{\sigma}_{a_h}^2 + \left(\frac{1}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{\sigma^2} \right) \widehat{\sigma}_{a_h}^2 \right) + \text{const.},
\end{aligned}$$

which is minimized when

$$\widehat{\sigma}_{a_h}^2 = \left(\frac{1}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{\sigma^2} \right)^{-1} = \frac{\sigma^2}{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2}. \quad (12.62)$$

Therefore, substituting Eqs. (12.61) and (12.62) into Eq. (12.60) gives the free energy with \widehat{a}_h and $\widehat{\sigma}_{a_h}^2$ already optimized:

$$\begin{aligned}
2\dot{F}_h^{\text{PB-A}} &= \min_{\widehat{a}_h, \widehat{\sigma}_{a_h}^2} 2F_h^{\text{PB-A}} \\
&= -M \log \widehat{\sigma}_{a_h}^2 + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + \frac{\widehat{b}_h^2}{c_{b_h}^2} - \frac{2\widehat{a}_h \widehat{b}_h \gamma_h}{\sigma^2} + \text{const.} \\
&= M \log(\widehat{b}_h^2 + \sigma^2/c_{a_h}^2) - \frac{\widehat{b}_h^2 \gamma_h^2}{\sigma^2(\widehat{b}_h^2 + \sigma^2/c_{a_h}^2)} + \frac{\widehat{b}_h^2}{c_{b_h}^2} + \text{const.} \\
&= M \log(\widehat{b}_h^2 + \sigma^2/c_{a_h}^2) + \left(\frac{\gamma_h^2}{\sigma^2} - \frac{\widehat{b}_h^2 \gamma_h^2}{\sigma^2(\widehat{b}_h^2 + \sigma^2/c_{a_h}^2)} \right) + \left(\frac{\widehat{b}_h^2}{c_{b_h}^2} + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right) + \text{const.} \\
&= M \log(\widehat{b}_h^2 + \sigma^2/c_{a_h}^2) + \frac{\gamma_h^2}{c_{a_h}^2} (\widehat{b}_h^2 + \sigma^2/c_{a_h}^2)^{-1} + \frac{1}{c_{b_h}^2} (\widehat{b}_h^2 + \sigma^2/c_{a_h}^2) + \text{const.}
\end{aligned} \quad (12.63)$$

In the second last equation, we added some constants so that the minimizer can be found by the following lemma:

Lemma 12.5 *The function*

$$f(x) = \xi_{\log} \log x + \xi_{-1} x^{-1} + \xi_1 x$$

of $x > 0$ for positive coefficients $\xi_{\log}, \xi_{-1}, \xi_1 > 0$ is strictly quasiconvex,² and minimized at

$$\widehat{x} = \frac{-\xi_{\log} + \sqrt{\xi_{\log}^2 + 4\xi_1\xi_{-1}}}{2\xi_1}.$$

Proof $f(x)$ is differentiable in $x > 0$, and its first derivative is

$$\begin{aligned} \frac{\partial f}{\partial x} &= \xi_{\log}x^{-1} - \xi_{-1}x^{-2} + \xi_1 \\ &= x^{-2}(\xi_1x^2 + \xi_{\log}x - \xi_{-1}) \\ &= x^{-2}\left(x + \frac{\xi_{\log} + \sqrt{\xi_{\log}^2 + 4\xi_1\xi_{-1}}}{2\xi_1}\right)\left(x - \frac{-\xi_{\log} + \sqrt{\xi_{\log}^2 + 4\xi_1\xi_{-1}}}{2\xi_1}\right). \end{aligned}$$

Since the first two factors are positive, we find that $f(x)$ is strictly decreasing for $0 < x < \widehat{x}$, and strictly increasing for $x > \widehat{x}$, which proves the lemma. \square

By applying Lemma 12.5 to Eq. (12.63) with $x = \widehat{b}_h^2 + \sigma^2/c_{a_h}^2$, we find that $\widehat{F}_h^{\text{PB-A}}$ is strictly quasiconvex and minimized when

$$\widehat{b}_h^2 + \sigma^2/c_{a_h}^2 = \frac{c_{b_h}^2 \left(-M + \sqrt{M^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)}{2}. \quad (12.64)$$

Since \widehat{b}_h^2 is nonnegative, the minimizer of the free energy (12.63) is given by

$$\widehat{b}_h^2 = \max \left\{ 0, \frac{c_{b_h}^2 \left(-\left(M + \frac{2\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right) + \sqrt{M^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)}{2} \right\}. \quad (12.65)$$

Apparently, Eq. (12.65) is positive when

$$\left(M + \frac{2\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right)^2 < M^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2},$$

which leads to the thresholding condition:

$$\gamma_h > \underline{\gamma}_h^{\text{PB-A}}.$$

By using Eqs. (12.61), (12.62), (12.65), and

$$\left(\widehat{b}_h^2 + \sigma^2/c_{a_h}^2 \right)^{-1} = \begin{cases} \frac{c_{a_h}^2 \left(M + \sqrt{M^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)}{2\gamma_h^2} & \text{if } \gamma_h > \underline{\gamma}_h^{\text{PB-A}}, \\ \frac{c_{a_h}^2}{\sigma^2} & \text{otherwise,} \end{cases} \quad (12.66)$$

² The definition of *quasiconvexity* is given in footnote 1 in Section 8.1.

derived from Eqs. (12.64) and (12.65), we obtain

$$\begin{aligned}\widehat{\gamma}_h^{\text{PB-A}} &\equiv \widehat{a}_h \widehat{b}_h = \frac{\widehat{b}_h^2 \gamma_h}{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2} \\&= \left(1 - \frac{\sigma^2/c_{a_h}^2}{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2}\right) \gamma_h \\&= \begin{cases} \left(1 - \frac{\sigma^2 \left(M + \sqrt{M^2 + 4 \frac{\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}}\right)}{2\gamma_h^2}\right) \gamma_h & \text{if } \gamma_h > \underline{\gamma}_h^{\text{PB-A}}, \\ 0 & \text{otherwise,} \end{cases} \\ \widehat{\delta}_h^{\text{PB-A}} &\equiv \frac{\widehat{a}_h}{\widehat{b}_h} = \frac{\gamma_h}{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2} \\&= \frac{c_{a_h}^2 \left(M + \sqrt{M^2 + 4 \frac{\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}}\right)}{2\gamma_h} \quad \text{for } \gamma_h > \underline{\gamma}_h^{\text{PB-A}}, \\ \widehat{\sigma}_{a_h}^2 &= \frac{\sigma^2}{\widehat{b}_h^2 + \sigma^2/c_{a_h}^2} \\&= \begin{cases} \frac{\sigma^2}{\underline{\gamma}_h^{\text{PB-A}}/\widehat{\delta}_h^{\text{PB-A}} + \sigma^2/c_{a_h}^2} & \text{if } \gamma_h > \underline{\gamma}_h^{\text{PB-A}}, \\ c_{a_h}^2 & \text{otherwise.} \end{cases}\end{aligned}$$

Thus, we have obtained the PB-A posterior (12.48) specified by Eqs. (12.49) through (12.53).

In exactly the same way, we can obtain the PB-B posterior (12.54) specified by Eqs. (12.55) through (12.59), by minimizing the free energy (12.32) for PB-B learning.

Finally, for the choice from the PB-A posterior and the PB-B posterior, we can easily prove the following lemma:

Lemma 12.6 *It holds that*

$$\begin{aligned}\min_{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2} F_h^{\text{PB-A}} &< \min_{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{b_h}^2} F_h^{\text{PB-B}} \quad \text{if} \quad L < M, \\ \min_{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2} F_h^{\text{PB-A}} &> \min_{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{b_h}^2} F_h^{\text{PB-B}} \quad \text{if} \quad L > M.\end{aligned}$$

Proof When comparing the PB-A free energy (12.31) and the PB-B free energy (12.32), the last terms are dominant since we assume that the negative entropy χ , defined by Eq. (12.33), of the one-dimensional pseudo-Dirac delta function is finite but arbitrarily large. Then, comparing the last terms, each

of which is proportional to the number of parameters point-estimated, of Eqs. (12.31) and (12.32) proves Lemma 12.6. \square

Combining Lemma 12.6 with the PB-A posterior and the PB-B posterior obtained before, we have Corollary 12.4. Theorem 12.3 is a direct consequence of Corollary 12.4. \square

Comparison between MAP, PB, and VB Solutions

Here we compare the MAP solution (Theorem 12.1), the PB solution (Theorem 12.3), and the VB solution (Theorem 6.7 in Chapter 6). For all methods, the solution is a shrinkage estimator applied to each singular value, i.e., in the following form for $\widehat{\gamma}_h \leq \gamma_h$:

$$\widehat{U} = \sum_{h=1}^H \widehat{\gamma}_h \omega_{b_h} \omega_{a_h}^\top, \quad \text{where} \quad \widehat{\gamma}_h = \begin{cases} \widehat{\gamma}_h & \text{if } \gamma_h \geq \underline{\gamma}_h, \\ 0 & \text{otherwise.} \end{cases} \quad (12.67)$$

When the prior is flat, i.e., $c_{a_h} c_{b_h} \rightarrow \infty$, the truncation threshold $\underline{\gamma}_h$ and the shrinkage factor $\check{\gamma}_h$ are simplified as

$$\lim_{c_{a_h} c_{b_h} \rightarrow \infty} \underline{\gamma}_h^{\text{MAP}} = 0, \quad \lim_{c_{a_h} c_{b_h} \rightarrow \infty} \check{\gamma}_h^{\text{MAP}} = \gamma_h, \quad (12.68)$$

$$\lim_{c_{a_h} c_{b_h} \rightarrow \infty} \underline{\gamma}_h^{\text{PB}} = \sigma \sqrt{\max(L, M)}, \quad \lim_{c_{a_h} c_{b_h} \rightarrow \infty} \check{\gamma}_h^{\text{PB}} = \left(1 - \frac{\max(L, M)\sigma^2}{\gamma_h}\right) \gamma_h, \quad (12.69)$$

$$\lim_{c_{a_h} c_{b_h} \rightarrow \infty} \underline{\gamma}_h^{\text{VB}} = \sigma \sqrt{\max(L, M)}, \quad \lim_{c_{a_h} c_{b_h} \rightarrow \infty} \check{\gamma}_h^{\text{VB}} = \left(1 - \frac{\max(L, M)\sigma^2}{\gamma_h}\right) \gamma_h, \quad (12.70)$$

and therefore the estimators $\widehat{\gamma}_h$ can be written as

$$\widehat{\gamma}_h^{\text{MAP}} = \gamma_h, \quad (12.71)$$

$$\widehat{\gamma}_h^{\text{PB}} = \max\left(0, \left(1 - \frac{\max(L, M)\sigma^2}{\gamma_h}\right) \gamma_h\right), \quad (12.72)$$

$$\widehat{\gamma}_h^{\text{VB}} = \max\left(0, \left(1 - \frac{\max(L, M)\sigma^2}{\gamma_h}\right) \gamma_h\right). \quad (12.73)$$

As expected, the MAP estimator (12.71) coincides with the maximum likelihood (ML) estimator when the prior is flat. On the other hand, the PB estimator (12.72) and the VB estimator (12.73) do not converge to the ML estimator. Interestingly, the PB estimator and the VB estimator coincide with each other, and they are in the form of the *positive-part James–Stein (PJS) estimator* (James and Stein, 1961; Efron and Morris, 1973) applied to each singular component (see Appendix A for a short introduction to the James–Stein estimator). The reason why the VB estimator is shrunken even with the flat prior was explained in Chapter 7 in terms of model-induced regularization

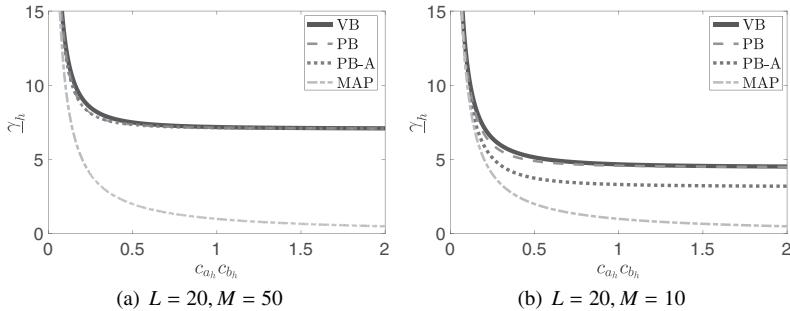


Figure 12.1 Truncation thresholds, γ_h^{VB} , γ_h^{PB} , $\gamma_h^{\text{PB-A}}$, and γ_h^{MAP} as functions of the product $c_{ah} c_{bh}$ of the prior covariances. The noise variance is set to $\sigma^2 = 1$.

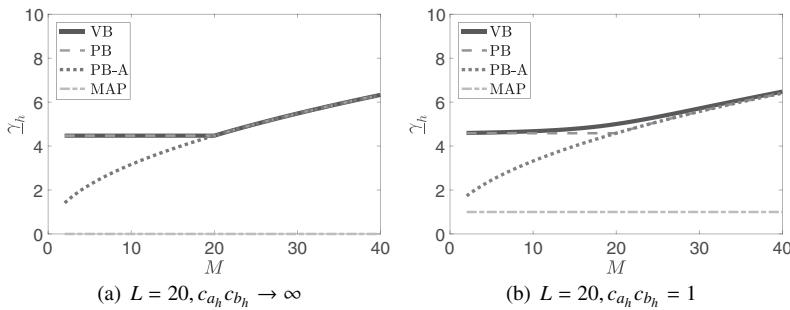


Figure 12.2 Truncation thresholds as functions of M .

(MIR) enhanced by phase transitions. Eq. (12.72) implies that PB learning—a cruder approximation to Bayesian learning—shares the same property as VB learning.

To investigate the dimensionality selection behavior, we depict the truncation thresholds of VB learning, PB learning, PB-A learning, and MAP learning as functions of the product $c_{a_h}c_{b_h}$ of the prior covariances in Figure 12.1. The left panel is for the case with $L = 20, M = 50$, and the right panel is for the case with $L = 20, M = 10$. PB-A learning corresponds to PB learning with the predetermined marginalized and the point-estimated spaces as in Tipping and Bishop (1999) and Chu and Ghahramani (2009), i.e., the matrix A is always marginalized out and \mathbf{B} is point-estimated regardless of the dimensionality. We see in Figure 12.1 that PB learning and VB learning show similar dimensionality selection behaviors, while PB-A learning behaves differently when $L > M$.

Figure 12.2 shows the truncation thresholds as functions of M for $L = 20$. With the flat prior $c_{a_k} c_{b_k} \rightarrow \infty$ (left panel), the PB and the VB solutions agree

with each other, as Eqs. (12.69) and (12.70) imply. The PB-A solution is also identical to them when $M \geq L$. However, its behavior changes at $M = L$: the truncation threshold of PB-A learning smoothly goes down as M decreases, while those of PB learning and VB learning make a sudden turn and become constant. The right panel is for the case with a nonflat prior ($c_{a_h}c_{b_h} = 1$), which shows similar tendency to the case with the flat prior.

A question is which behavior is more desirable, a sudden turn in the threshold curve in VB/PB learning, or the smooth behavior in PB-A learning? We argue that the behavior of VB/PB learning is more desirable for the following reason. Let us consider the case where no *true* signal exists, i.e., the true rank is $H^* = 0$. In this case, we merely observe pure noise, $\mathbf{V} = \mathcal{E}$, and the average of the squared singular values of \mathbf{V} over all components is given by

$$\frac{\langle \text{tr}(\mathcal{E}\mathcal{E}^\top) \rangle_{\text{MGauss}_{L,M}(\mathcal{E}; \mathbf{0}_{L,M}, \sigma^2 I_L \otimes I_M)}}{\min(L, M)} = \sigma^2 \max(L, M). \quad (12.74)$$

Comparing Eq. (12.74) with Eqs. (12.70) and (12.69), we find that VB learning and PB learning always discard the components with singular values no greater than the average noise contribution (note here that Eqs. (12.70) and (12.69) give the thresholds for the flat prior $c_{a_h}c_{b_h} \rightarrow \infty$, and the thresholds increase as $c_{a_h}c_{b_h}$ decreases). The sudden turn in the threshold curve actually follows the behavior of the average noise contribution (12.74) to the singular values. On the other hand, PB-A learning does not necessarily discard such noise-dominant components, and can strongly overfit the noise when $L \gg M$.

Figure 12.3 shows the estimators $\hat{\gamma}_h$ by VB learning, PB learning, PB-A learning, and MAP learning for $c_{a_h}c_{b_h} = 1$, as functions of the observed

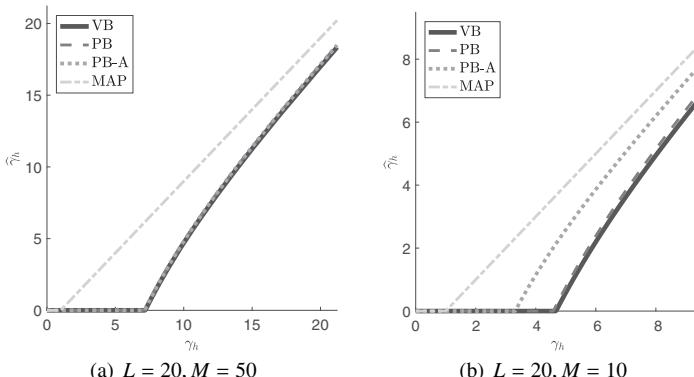


Figure 12.3 Behavior of VB, PB, PB-A, and MAP estimators (the vertical axis) for $c_{a_h}c_{b_h} = 1$, when the singular value γ_h (the horizontal axis) is observed. The noise variance is set to $\sigma^2 = 1$.

singular value γ_h . We can see that the PB estimator behaves similarly to the VB estimator, while the MAP estimator behaves significantly differently. The right panel shows that PB-A learning also behaves differently from VB learning when $L > M$, which implies that the choice between the PB-A posterior and the PB-B posterior based on the free energy is essential to accurately approximate VB learning.

Actually, the coincidence between the VB solution (12.70) and the PB solution (12.69) with the flat prior can be seen as a natural consequence from the similarity in the posterior shape. From Theorem 6.7 and Corollary 6.8, we can derive the following corollary:

Corollary 12.7 *Assume that, when we make the prior flat $c_{a_h}c_{b_h} \rightarrow \infty$, c_{a_h} and c_{b_h} go to infinity in the same order, i.e., $c_{a_h}/c_{b_h} = \Theta(1)$.³ Then, the following hold for the variances of the VB posterior (6.40): when $L < M$,*

$$\lim_{c_{a_h}c_{b_h} \rightarrow \infty} \widehat{\sigma}_{a_h}^2 = \infty, \quad \lim_{c_{a_h}c_{b_h} \rightarrow \infty} \widehat{\sigma}_{b_h}^2 = 0, \quad (12.75)$$

and when $L > M$,

$$\lim_{c_{a_h}c_{b_h} \rightarrow \infty} \widehat{\sigma}_{a_h}^2 = 0, \quad \lim_{c_{a_h}c_{b_h} \rightarrow \infty} \widehat{\sigma}_{b_h}^2 = \infty. \quad (12.76)$$

Proof Assume first that $L < M$. When $\gamma_h > \underline{\gamma}_h^{\text{VB}}$, Eq. (6.52) gives $\lim_{c_{a_h}c_{b_h} \rightarrow \infty} \widehat{\delta}_h^{\text{VB}} = \infty$, and therefore Eq. (6.51) gives Eq. (12.75). When $\gamma_h \leq \underline{\gamma}_h^{\text{VB}}$, Eq. (6.54) gives $\widehat{\zeta}_h^{\text{VB}} = \sigma^2/M - \Theta(c_{a_h}^{-2}c_{b_h}^{-2})$ as $c_{a_h}c_{b_h} \rightarrow \infty$, and therefore Eq. (6.53) gives Eq. (12.75).

When $L > M$, Theorem 6.7 and Corollary 6.8 hold for $\mathbf{V} \leftarrow \mathbf{V}^\top$, meaning that the VB posterior is obtained by exchanging the variational parameters for \mathbf{A} and those for \mathbf{B} . Thus, we obtain Eq. (12.76), and complete the proof. \square

Corollary 12.7 implies that, with the flat prior $c_{a_h}c_{b_h} \rightarrow \infty$, the shape of the VB posterior is similar to the shape of the PB posterior: they extend in the space of \mathbf{A} when $M > L$, and extend in the space of \mathbf{B} when $M < L$. Therefore, it is no wonder that the solutions coincide with each other.

Global Empirical MAP and Empirical PB Solutions

Next, we investigate the empirical Bayesian variants of MAP learning and PB learning, where the hyperparameters \mathbf{C}_A and \mathbf{C}_B are also estimated from observations. Note that the noise variance σ^2 is still considered as a fixed constant (noise variance estimation will be discussed in Section 12.1.6).

³ $\Theta(f(x))$ is a positive function such that $\limsup_{x \rightarrow \infty} |\Theta(f(x))/f(x)| < \infty$ and $\liminf_{x \rightarrow \infty} |\Theta(f(x))/f(x)| > 0$.

Let us first consider the MAP free energy (12.30):

$$2F_h^{\text{MAP}} = M \log c_{a_h}^2 + L \log c_{b_h}^2 + \frac{\widehat{a}_h^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{c_{b_h}^2} + \frac{-2\widehat{a}_h\widehat{b}_h\gamma_h + \widehat{a}_h^2\widehat{b}_h^2}{\sigma^2} - (L+M) + (L+M)\chi.$$

We can make the MAP free energy arbitrarily small, i.e., $F_h^{\text{MAP}} \rightarrow -\infty$ by setting $c_{a_h}^2, c_{b_h}^2 \rightarrow +0$ with the variational parameters set to the corresponding solution, i.e., $\widehat{a}_h = \widehat{b}_h = 0$ (see Corollary 12.2). Therefore, the global solution of *empirical MAP (EMAP) learning* is given by

$$\widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad c_{a_h}^2 \rightarrow +0, \quad c_{b_h}^2 \rightarrow +0, \quad \text{for } h = 1, \dots, H,$$

which results in the following theorem:

Theorem 12.8 *The global solution of EMAP learning is $\widehat{\gamma}_h^{\text{EMAP}} (\equiv \widehat{a}_h\widehat{b}_h) = 0$ for all $h = 1, \dots, H$, regardless of observations.*

The same happens in *empirical PB (EPB) learning*. The PB-A free energy (12.31),

$$2F_h^{\text{PB-A}} = M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log c_{b_h}^2 + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2}{c_{b_h}^2} + \frac{-2\widehat{a}_h\widehat{b}_h\gamma_h + (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2)\widehat{b}_h^2}{\sigma^2} - (L+M) + L\chi,$$

can be arbitrarily small, i.e., $F_h^{\text{PB-A}} \rightarrow -\infty$ by setting $c_{a_h}^2, c_{b_h}^2 \rightarrow +0$ with the variational parameters set to the corresponding solution $\widehat{a}_h = \widehat{b}_h = 0, \widehat{\sigma}_{a_h}^2 = c_{a_h}^2$ (see Corollary 12.4). Also, the PB-B free energy (12.32),

$$2F_h^{\text{PB-B}} = M \log c_{a_h}^2 + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{\widehat{a}_h^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2}{c_{b_h}^2} + \frac{-2\widehat{a}_h\widehat{b}_h\gamma_h + \widehat{a}_h^2(\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2)}{\sigma^2} - (L+M) + M\chi,$$

can be arbitrarily small, i.e., $F_h^{\text{PB-B}} \rightarrow -\infty$ by setting $c_{a_h}^2, c_{b_h}^2 \rightarrow +0$ with the variational parameters set to the corresponding solution $\widehat{a}_h = \widehat{b}_h = 0, \widehat{\sigma}_{b_h}^2 = c_{b_h}^2$. Thus, we have the following theorem:

Theorem 12.9 *The global solution of EPB learning is $\widehat{\gamma}_h^{\text{EPB}} (\equiv \widehat{a}_h\widehat{b}_h) = 0$ for all $h = 1, \dots, H$, regardless of observations.*

Theorems 12.8 and 12.9 imply that empirical Bayesian variants of MAP learning and PB learning give useless trivial estimators. This happens because the posterior variances of the parameters to be point-estimated are fixed to a small value, so that the posteriors form the pseudo-Dirac delta functions. In VB learning, if we set $c_{a_h}c_{b_h}$ to a small value, the posterior variances, $\widehat{\sigma}_{a_h}^2$ and $\widehat{\sigma}_{b_h}^2$,

get small accordingly, so that the third and the fourth terms in Eq. (12.27) do not diverge to $+\infty$. As a result, the first and the second terms in Eq. (12.27) remain finite. On the other hand, in MAP learning and PB learning, at least one of the posterior variances, $\widehat{\sigma}_{a_h}^2$ and $\widehat{\sigma}_{b_h}^2$, is treated as a constant and cannot be adjusted to the corresponding prior covariance when it is set to be small. This makes the free energy lower-unbounded. Actually, if we lower-bound the prior covariances as $c_{a_h}^2, c_{b_h}^2 \geq \varepsilon^2$ with the same ε^2 as the one we used for defining the variances (12.28) and (12.29) of the pseudo-Dirac delta functions and their entropy (12.33), the MAP and the PB free energies, F_h^{MAP} , $F_h^{\text{PB-A}}$, and $F_h^{\text{PB-B}}$, are also lower-bounded by zero, as the VB free energy, F_h^{VB} .

12.1.3 Local Solutions

The analysis in Section 12.1.2 might seem contradictory with the reported results in Mørup and Hansen (2009), where EMAP showed good performance with the ARD property in TF—since the free energies in MF and TF are similar to each other, they should share the same issue of the lower-unboundedness. In the following, we elucidate that this apparent contradiction is because of the local solutions in EMAP learning and EPB learning that behave similarly to the nontrivial positive solution of EVB learning. Actually, EMAP learning and EPB learning can behave similarly to EVB learning when the free energy is minimized by local search.

Local EMAP and EPB Solutions

Here we conduct more detailed analysis of the free energies for EMAP learning and EPB learning, and clarify the behavior of their local minima. To make the free energy always comparable (finite), we slightly modify the problem. Specifically, we solve the following problem:

$$\begin{aligned} \text{Given } \quad & \sigma^2 \in \mathbb{R}_{++}, \\ & \min_{r, \{c_{a_h}^2, c_{b_h}^2\}_{h=1}^H} F, \\ \text{s.t. } \quad & c_{a_h} c_{b_h} \geq \varepsilon^2, \quad c_{a_h}/c_{b_h} = 1 \quad \text{for } h = 1, \dots, H, \\ \text{and } & \begin{cases} r(\mathbf{A}, \mathbf{B}) = \delta(\mathbf{A}; \widehat{\mathbf{A}})\delta(\mathbf{B}; \widehat{\mathbf{B}}) & (\text{for EMAP learning}), \\ r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})\delta(\mathbf{B}; \widehat{\mathbf{B}}) & (\text{for EPB-A learning}), \\ r(\mathbf{A}, \mathbf{B}) = \delta(\mathbf{A}; \widehat{\mathbf{A}})r_B(\mathbf{B}) & (\text{for EPB-B learning}), \\ r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}) & (\text{for EVB learning}), \end{cases} \end{aligned} \tag{12.77}$$

where the free energy F is defined by Eq. (12.6), and the pseudo-Dirac delta function is defined as Gaussian with an arbitrarily small but positive variance $\varepsilon^2 > 0$:

$$\delta(\mathbf{A}; \widehat{\mathbf{A}}) = \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \varepsilon^2 \mathbf{I}_M \otimes \mathbf{I}_H) \propto \exp\left(-\frac{\|\mathbf{A} - \widehat{\mathbf{A}}\|_{\text{Fro}}^2}{2\varepsilon^2}\right),$$

$$\delta(\mathbf{B}; \widehat{\mathbf{B}}) = \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \varepsilon^2 \mathbf{I}_L \otimes \mathbf{I}_H) \propto \exp\left(-\frac{\|\mathbf{B} - \widehat{\mathbf{B}}\|_{\text{Fro}}^2}{2\varepsilon^2}\right).$$

Note that, in Eq. (12.77), we lower-bounded the product $c_{a_h} c_{b_h}$ of the prior covariances and fixed the ratio c_{a_h}/c_{b_h} . We added the constraint for EVB learning for comparison.

Following the discussion in Section 12.1.1, we can write the posterior as

$$r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}), \quad \text{where}$$

$$r_A(\mathbf{A}) = \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \mathbf{I}_M \otimes \widehat{\Sigma}_A) \propto \exp\left(-\frac{\text{tr}\left((\mathbf{A} - \widehat{\mathbf{A}})\widehat{\Sigma}_A^{-1}(\mathbf{A} - \widehat{\mathbf{A}})^T\right)}{2}\right),$$

$$r_B(\mathbf{B}) = \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \mathbf{I}_L \otimes \widehat{\Sigma}_B) \propto \exp\left(-\frac{\text{tr}\left((\mathbf{B} - \widehat{\mathbf{B}})\widehat{\Sigma}_B^{-1}(\mathbf{B} - \widehat{\mathbf{B}})^T\right)}{2}\right),$$

for

$$\begin{aligned} \widehat{\mathbf{A}} &= (\widehat{a}_1 \omega_{a_1}, \dots, \widehat{a}_H \omega_{a_H}), \\ \widehat{\mathbf{B}} &= (\widehat{b}_1 \omega_{b_1}, \dots, \widehat{b}_H \omega_{b_H}), \\ \widehat{\Sigma}_A &= \mathbf{Diag}(\widehat{\sigma}_{a_1}^2, \dots, \widehat{\sigma}_{a_H}^2), \\ \widehat{\Sigma}_B &= \mathbf{Diag}(\widehat{\sigma}_{b_1}^2, \dots, \widehat{\sigma}_{b_H}^2), \end{aligned}$$

and the variational parameters $\{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2\}_{h=1}^H$ are the solution of the following problem:

$$\begin{aligned} \text{Given } \sigma^2 \in \mathbb{R}_{++}, \quad & \min_{\{\widehat{a}_h, \widehat{b}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2, c_{a_h}^2, c_{b_h}^2\}_{h=1}^H} F, \\ \text{s.t. } \widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \quad c_{a_h} c_{b_h} & \geq \varepsilon^2, \quad c_{a_h}/c_{b_h} = 1, \end{aligned} \tag{12.78}$$

$$\text{and } \begin{cases} \widehat{\sigma}_{a_h}^2 = \varepsilon^2, & \widehat{\sigma}_{b_h}^2 = \varepsilon^2 \quad (\text{for EMAP learning}), \\ \widehat{\sigma}_{a_h}^2 \geq \varepsilon^2, & \widehat{\sigma}_{b_h}^2 = \varepsilon^2 \quad (\text{for EPB-A learning}), \\ \widehat{\sigma}_{a_h}^2 = \varepsilon^2, & \widehat{\sigma}_{b_h}^2 \geq \varepsilon^2 \quad (\text{for EPB-B learning}), \\ \widehat{\sigma}_{a_h}^2 \geq \varepsilon^2, & \widehat{\sigma}_{b_h}^2 \geq \varepsilon^2 \quad (\text{for EVB learning}), \end{cases} \tag{12.79}$$

$$\text{for } h = 1, \dots, H,$$

where the free energy F is explicitly written by Eqs. (12.26) and (12.27), that is,

$$2F = LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^{\min(L,M)} \gamma_h^2}{\sigma^2} + \sum_{h=1}^H 2F_h, \quad (12.80)$$

where

$$\begin{aligned} 2F_h &= M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2}{c_{b_h}^2} \\ &\quad - (L+M) + \frac{-2\widehat{a}_h\widehat{b}_h\gamma_h + (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2)(\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2)}{\sigma^2}. \end{aligned} \quad (12.81)$$

By substituting the MAP solution (Corollary 12.2) and the PB solution (Corollary 12.4), respectively, into Eq. (12.81), we can write the free energy as a function of the product $c_{a_h}c_{b_h}$ of the prior covariances. We have the following lemmas (the proofs are given in Sections 12.1.4 and 12.1.5, respectively):

Lemma 12.10 *In EMAP learning, the free energy (12.81) can be written as a function of $c_{a_h}c_{b_h}$ as follows:*

$$\begin{aligned} 2\dot{F}_h^{\text{MAP}} &= \min_{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2 = \widehat{\sigma}_{b_h}^2 = \varepsilon^2} 2F_h \\ &= \begin{cases} (L+M) \left(\log c_{a_h}c_{b_h} + \frac{\varepsilon^2}{c_{a_h}c_{b_h}} - 1 + \chi \right) & \text{for } \varepsilon^2 \leq c_{a_h}c_{b_h} \leq \frac{\sigma^2}{\gamma_h}, \\ (L+M) \left(\log c_{a_h}c_{b_h} - 1 + \chi \right) - \sigma^{-2} \left(\gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}} \right)^2 & \text{for } c_{a_h}c_{b_h} > \frac{\sigma^2}{\gamma_h}. \end{cases} \end{aligned} \quad (12.82)$$

Lemma 12.11 *In EPB learning, the free energy (12.81) can be written as a function of $c_{a_h}c_{b_h}$ as follows: if $\gamma_h > \sigma \sqrt{\max(L, M)}$,*

$$\begin{aligned} 2\dot{F}_h^{\text{PB}} &= \min \{2\dot{F}_h^{\text{PB-A}}, 2\dot{F}_h^{\text{PB-B}}\} \\ &= \min \left\{ \min_{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2 \geq \varepsilon^2, \widehat{\sigma}_{b_h}^2 = \varepsilon^2} 2F_h, \min_{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2 = \varepsilon^2, \widehat{\sigma}_{b_h}^2 \geq \varepsilon^2} 2F_h \right\} \\ &= \begin{cases} \min(L, M) \left(\log c_{a_h}c_{b_h} + \frac{\varepsilon^2}{c_{a_h}c_{b_h}} - 1 + \chi \right) & \text{for } \varepsilon^2 \leq c_{a_h}c_{b_h} \leq \frac{\sigma^2}{\sqrt{\gamma_h^2 - \max(L, M)\sigma^2}}, \\ \frac{\min(L, M) + 2\max(L, M)}{2} \log c_{a_h}^2 c_{b_h}^2 + \sqrt{\max(L, M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2} - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}} \\ \quad + \max(L, M) \log \left(-\max(L, M) + \sqrt{\max(L, M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) \\ \quad - \frac{\gamma_h^2}{\sigma^2} - \max(L, M) \log(2\sigma^2) + \min(L, M)(\chi - 1) & \text{for } c_{a_h}c_{b_h} > \frac{\sigma^2}{\sqrt{\gamma_h^2 - \max(L, M)\sigma^2}}, \end{cases} \end{aligned} \quad (12.83)$$

and otherwise,

$$2\hat{F}_h^{\text{PB}} = \min(L, M) \left(\log c_{a_h} c_{b_h} + \frac{\varepsilon^2}{c_{a_h} c_{b_h}} - 1 + \chi \right). \quad (12.84)$$

By minimizing the EMAP free energy (12.82) and the EPB free energy (12.83), respectively, with respect to the product $c_{a_h} c_{b_h}$ of the prior covariances, we obtain the following theorems (the proofs are given also in Sections 12.1.4 and 12.1.5, respectively):

Theorem 12.12 *In EMAP learning, the free energy (12.81) has the global minimum such that*

$$\widehat{\gamma}_h^{\text{EMAP}} \left(\equiv \widehat{a}_h \widehat{b}_h \right) = 0.$$

It has a nontrivial local minimum such that

$$\widehat{\gamma}_h^{\text{local-EMAP}} \left(\equiv \widehat{a}_h \widehat{b}_h \right) = \check{\gamma}_h^{\text{local-EMAP}} \quad \text{if and only if} \quad \gamma_h > \underline{\gamma}^{\text{local-EMAP}},$$

where

$$\underline{\gamma}^{\text{local-EMAP}} = \sigma \sqrt{2(L + M)}, \quad (12.85)$$

$$\check{\gamma}_h^{\text{local-EMAP}} = \frac{1}{2} \left(\gamma_h + \sqrt{\gamma_h^2 - 2\sigma^2(L + M)} \right). \quad (12.86)$$

Theorem 12.13 *In EPB learning, the free energy (12.81) has the global minimum such that*

$$\widehat{\gamma}_h^{\text{EPB}} \left(\equiv \widehat{a}_h \widehat{b}_h \right) = 0.$$

It has a non-trivial local minimum such that

$$\widehat{\gamma}_h^{\text{local-EPB}} \left(\equiv \widehat{a}_h \widehat{b}_h \right) = \check{\gamma}_h^{\text{local-EPB}} \quad \text{if and only if} \quad \gamma_h > \underline{\gamma}^{\text{local-EPB}},$$

where

$$\underline{\gamma}^{\text{local-EPB}} = \sigma \sqrt{L + M + \sqrt{2LM + \min(L, M)^2}}, \quad (12.87)$$

$$\check{\gamma}_h^{\text{local-EPB}} = \frac{\gamma_h}{2} \left(1 + \frac{-\max(L, M)\sigma^2 + \sqrt{\gamma_h^4 - 2(L+M)\sigma^2\gamma_h^2 + \min(L, M)^2\sigma^4}}{\gamma_h^2} \right). \quad (12.88)$$

Figure 12.4 shows the free energy (normalized by LM) as a function of $c_{a_h} c_{b_h}$ for EMAP learning (given in Lemma 12.10), EPB learning (given in Lemma 12.11), and EVB learning, defined by

$$2\hat{F}_h^{\text{VB}} = \min_{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2 \geq \varepsilon^2} 2F_h.$$

For EMAP learning and EPB learning, we ignored some constants (e.g., the entropy terms proportional to χ) to make the *shapes* of the free energies

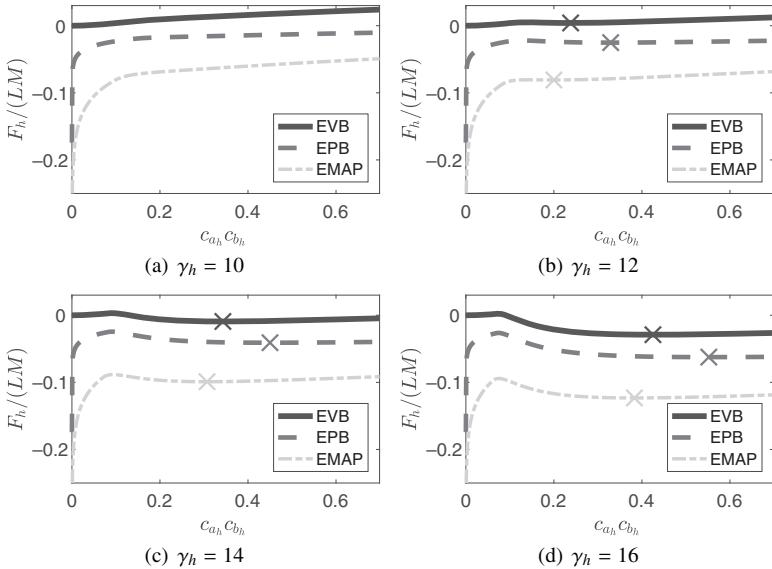


Figure 12.4 Free energy dependence on $c_{ah}c_{bh}$, where $L = 20, M = 50$. Crosses indicate nontrivial local minima.

comparable. We can see deep pits at $c_{ah}c_{bh} \rightarrow +0$ in EMAP and EPB free energies, which correspond to the global solutions. However, we also see nontrivial local minima, which behave similarly to the nontrivial local solution for VB learning. Namely, nontrivial local minima of EMAP, EPB, and EVB free energies appear at locations similar to each other when the observed singular value γ_h exceeds the thresholds given by Eqs. (12.85), (12.87), and (6.127), respectively.

The deep pit at $c_{ah}c_{bh} \rightarrow +0$ is essential when we stick to the global solution. The VB free energy does not have such a pit, which enables consistent inference based on the free energy minimization principle. However, as long as we rely on local search, the pit at the origin is not essential in practice. Assume that a nontrivial local minimum exists, and we perform local search only once. Then, whether local search for EMAP learning or EPB learning converges to the trivial global solution or the nontrivial local solution simply depends on the initialization. Note that the same applies also to EVB learning, for which local search is not guaranteed to converge to the global solution. This is because of the multimodality of the VB free energy, which can be seen in Figure 12.4.

One might wonder if some hyperpriors on c_{ah}^2 and c_{bh}^2 could fill the deep pits at $c_{ah}c_{bh} \rightarrow +0$ in the EMAP and the EPB free energies, so that the nontrivial

local solutions are global when some reasonable conditions hold. However, when we rely on the ARD property for model selection, hyperpriors should be almost noninformative. With such an almost noninformative hyperprior, e.g., the inverse-Gamma, $p(c_{a_h}^2, c_{b_h}^2) \propto (c_{a_h}^2 c_{b_h}^2)^{1.001} + 0.001/(c_{a_h}^2 c_{b_h}^2)$, which was used in Bishop (1999b), deep pits still exist very close to the origin, which keep the global EMAP and EPB estimators useless.

Comparison between Local-EMAP, Local-EPB, and EVB Solutions

Let us observe the behavior of local solutions. We define the *local-EMAP estimator* and the *local-EPB estimator*, respectively, by

$$\widehat{\boldsymbol{U}}^{\text{local-EMAP}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{local-EMAP}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top,$$

where $\widehat{\gamma}_h^{\text{local-EMAP}} = \begin{cases} \check{\gamma}_h^{\text{local-EMAP}} & \text{if } \gamma_h \geq \underline{\gamma}^{\text{local-EMAP}}, \\ 0 & \text{otherwise,} \end{cases}$ (12.89)

$$\widehat{\boldsymbol{U}}^{\text{local-EPB}} = \sum_{h=1}^H \widehat{\gamma}_h^{\text{local-EPB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top,$$

where $\widehat{\gamma}_h^{\text{local-EPB}} = \begin{cases} \check{\gamma}_h^{\text{local-EPB}} & \text{if } \gamma_h \geq \underline{\gamma}^{\text{local-EPB}}, \\ 0 & \text{otherwise,} \end{cases}$ (12.90)

following the definition of the local-EVB estimator (6.131) in Chapter 6. In the following, we assume that local search algorithms for EMAP learning and EPB learning find these solutions.

Define the normalized (by the average noise contribution (12.74)) singular values:

$$\gamma'_h = \frac{\gamma_h}{\sqrt{\max(L, M)\sigma^2}}.$$

We also define normalized versions of the estimator, the truncation threshold, and the shrinkage factor as

$$\widehat{\gamma}'_h = \frac{\widehat{\gamma}_h}{\sqrt{\max(L, M)\sigma^2}}, \gamma'_h = \frac{\gamma_h}{\sqrt{\max(L, M)\sigma^2}}, \check{\gamma}'_h = \frac{\check{\gamma}_h}{\sqrt{\max(L, M)\sigma^2}}. \quad (12.91)$$

Then the normalized truncation thresholds and the normalized shrinkage factors can be written as functions of $\alpha = \min(L, M)/\max(L, M)$ as follows:

$$\underline{\gamma}'^{\text{EVB}} = \sigma \sqrt{1 + \alpha + \sqrt{\alpha} \left(\kappa + \frac{1}{\kappa} \right)}, \quad (12.92)$$

$$\check{\gamma}'^{\text{EVB}} = \frac{\gamma'_h}{2} \left(1 - \frac{(1+\alpha)\sigma^2}{\gamma'^2_h} + \sqrt{\left(1 - \frac{(1+\alpha)\sigma^2}{\gamma'^2_h} \right)^2 - \frac{4\alpha\sigma^4}{\gamma'^4_h}} \right), \quad (12.93)$$

$$\underline{\gamma}'^{\text{local-EPB}} = \sigma \sqrt{1 + \alpha + \sqrt{2\alpha + \alpha^2}}, \quad (12.94)$$

$$\check{\gamma}_h'^{\text{local-EPB}} = \frac{\gamma_h'}{2} \left(1 + \frac{-\sigma^2 + \sqrt{\gamma_h'^4 - 2(1+\alpha)\sigma^2\gamma_h'^2 + \sigma^4}}{\gamma_h'^2} \right), \quad (12.95)$$

$$\underline{\gamma}'^{\text{local-EMAP}} = \sigma \sqrt{2(1 + \alpha)}, \quad (12.96)$$

$$\check{\gamma}_h'^{\text{local-EMAP}} = \frac{1}{2} \left(\gamma_h' + \sqrt{\gamma_h'^2 - 2\sigma^2(1 + \alpha)} \right). \quad (12.97)$$

Note that $\underline{\kappa}$ is also a function of α .

Figure 12.5 compares the normalized versions of the (global) EVB estimator $\check{\gamma}_h^{\text{EVB}}$, the local-EPB estimator $\check{\gamma}_h'^{\text{local-EPB}}$, and the local-EMAP estimator $\check{\gamma}_h'^{\text{local-EMAP}}$. We can observe similar behaviors of those three empirical Bayesian estimators. This is in contrast to the nonempirical Bayesian estimators shown in Figure 12.3, where the PB estimator behaves similarly to the VB estimator, while the MAP estimator behaves differently.

Figure 12.6 compares the normalized versions of the EVB truncation threshold (12.92), the local-EPB truncation threshold (12.94), and the local-EMAP truncation threshold (12.96). We can see that those thresholds behave similarly. However, we can find an essential difference of the local-EPB threshold from the EVB and the local-EMAP thresholds: it holds that, for any α ,

$$\underline{\gamma}'^{\text{local-EPB}} < \bar{\gamma}'^{\text{MPUL}} \leq \underline{\gamma}'^{\text{EVB}}, \underline{\gamma}'^{\text{local-EMAP}}, \quad (12.98)$$

$$\text{where } \bar{\gamma}'^{\text{MPUL}} = \frac{\bar{\gamma}^{\text{MPUL}}}{\sqrt{\max(L, M)\sigma^2}} = (1 + \sqrt{\alpha})$$

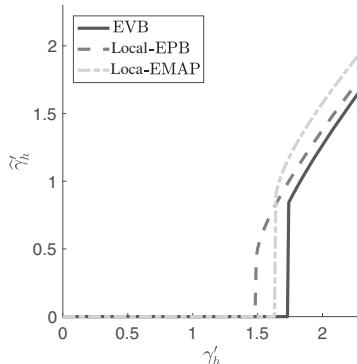


Figure 12.5 Behavior of (global) EVB, the local-EPB, and the local-EMAP estimators for $\alpha = \min(L, M)/\max(L, M) = 1/3$.

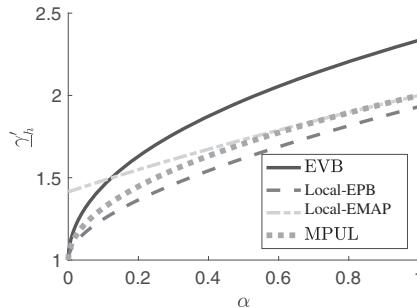


Figure 12.6 Truncation thresholds.

is the normalized version of the Marčenko–Pastur upper limit (MPUL) (Eq. (8.41) in Chapter 8), which is also shown in Figure 12.6.

As discussed in Section 8.4.1, the MPUL is the largest singular value of an $L \times M$ zero-mean independent random matrix in the *large-scale limit* where the matrix size (L, M) goes to infinity with fixed ratio $\alpha = \min(L, M)/\max(L, M)$. In other words, the MPUL corresponds to the minimum observed singular value detectable (or distinguishable from noise) by any dimensionality reduction method. The inequalities (12.98) say that local-EPB threshold is always smaller than the MPUL, while the EVB threshold and the local-EMAP threshold are never smaller than the MPUL. This implies that, for a large-scale observed matrix, the EVB estimator and the local-EMAP estimator discard the singular components dominated by noise, while the local-EPB estimator retains some of them.

12.1.4 Proofs of Lemma 12.10 and Theorem 12.12

By substituting the MAP solution, given by Corollary 12.2, we can write the free energy (12.81) as follows: for $\varepsilon^2 \leq c_{a_h} c_{b_h} \leq \frac{\sigma^2}{\gamma_h}$,

$$\begin{aligned}
 2\hat{F}_h^{\text{MAP}} &= \min_{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2 = \widehat{\sigma}_{b_h}^2 = \varepsilon^2} 2F_h \\
 &= M \log c_{a_h}^2 + L \log c_{b_h}^2 \\
 &\quad + \left(\frac{M}{c_{a_h}^2} + \frac{L}{c_{b_h}^2} \right) \varepsilon^2 - (L + M) + (L + M)\chi \\
 &= M \log c_{a_h}^2 + L \log c_{b_h}^2 + \left(\frac{M}{c_{a_h}^2} + \frac{L}{c_{b_h}^2} \right) \varepsilon^2 \\
 &\quad - (L + M) + (L + M)\chi,
 \end{aligned} \tag{12.99}$$

and for $c_{a_h}c_{b_h} > \frac{\sigma^2}{\gamma_h}$,

$$\begin{aligned} 2\hat{F}_h^{\text{MAP}} &= \min_{\hat{a}_h, \hat{b}_h \in \mathbb{R}, \hat{\sigma}_{a_h}^2 = \hat{\sigma}_{b_h}^2 = \varepsilon^2} 2F_h \\ &= M \log c_{a_h}^2 + L \log c_{b_h}^2 + \left(\gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}} \right) \left(\frac{2}{c_{a_h}c_{b_h}} + \frac{-2\gamma_h + \gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}}}{\sigma^2} \right) \\ &\quad + \left(\frac{M}{c_{a_h}^2} + \frac{L}{c_{b_h}^2} \right) \varepsilon^2 - (L + M) + (L + M)\chi \\ &= M \log c_{a_h}^2 + L \log c_{b_h}^2 - \sigma^{-2} \left(\gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}} \right)^2 \\ &\quad - (L + M) + (L + M)\chi. \end{aligned} \quad (12.100)$$

In the second-to-last equation in Eq. (12.100), we ignored the fourth term because $c_{a_h}c_{b_h} > \frac{\sigma^2}{\gamma_h}$ implies $c_{a_h}^2, c_{b_h}^2 \gg \varepsilon^2$ (with an arbitrarily high probability depending on ε^2). By fixing the ratio to $c_{a_h}/c_{b_h} = 1$, we obtain Eq. (12.82), which proves Lemma 12.10.

Now we minimize the free energy (12.82) with respect to $c_{a_h}c_{b_h}$, and find nontrivial local solutions. The free energy is continuous in the domain $\varepsilon^2 \leq c_{a_h}c_{b_h} < \infty$, and differentiable except at $c_{a_h}c_{b_h} = \frac{\sigma^2}{\gamma_h}$. The derivative is given by

$$\begin{aligned} \frac{\partial 2\hat{F}_h^{\text{MAP}}}{\partial (c_{a_h}c_{b_h})} &= \begin{cases} (L + M) \left(\frac{1}{c_{a_h}c_{b_h}} - \frac{\varepsilon^2}{c_{a_h}^2c_{b_h}^2} \right) & \text{for } \varepsilon^2 \leq c_{a_h}c_{b_h} \leq \frac{\sigma^2}{\gamma_h} \\ \left(\frac{L+M}{c_{a_h}c_{b_h}} + 2\sigma^{-2} \left(\gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}} \right) \frac{\sigma^2}{c_{a_h}^2c_{b_h}^2} \right) & \text{for } c_{a_h}c_{b_h} > \frac{\sigma^2}{\gamma_h} \end{cases} \\ &= \begin{cases} \frac{L+M}{c_{a_h}^2c_{b_h}^2} (c_{a_h}c_{b_h} - \varepsilon^2) & \text{for } \varepsilon^2 \leq c_{a_h}c_{b_h} \leq \frac{\sigma^2}{\gamma_h}, \\ \frac{1}{c_{a_h}^3c_{b_h}^3} ((L + M)c_{a_h}^2c_{b_h}^2 + 2\gamma_h c_{a_h}c_{b_h} - 2\sigma^2) & \text{for } c_{a_h}c_{b_h} > \frac{\sigma^2}{\gamma_h}. \end{cases} \end{aligned} \quad (12.101)$$

Eq. (12.101) implies that the free energy \hat{F}_h^{MAP} is increasing for $\varepsilon^2 \leq c_{a_h}c_{b_h} \leq \frac{\sigma^2}{\gamma_h}$, and that it is increasing at $c_{a_h}c_{b_h} = \frac{\sigma^2}{\gamma_h}$ and at $c_{a_h}c_{b_h} \rightarrow +\infty$.

In the region of $\frac{\sigma^2}{\gamma_h} < c_{a_h}c_{b_h} < +\infty$, the free energy has stationary points if and only if

$$\gamma_h \geq \sigma \sqrt{2(L + M)} \left(\equiv \underline{\gamma}^{\text{local-EMAP}} \right), \quad (12.102)$$

because the derivative can be factorized (with real factors if and only if the condition (12.102) holds) as

$$\frac{\partial 2\hat{F}_h^{\text{MAP}}}{\partial (c_{a_h}c_{b_h})} = \frac{L + M}{c_{a_h}^3c_{b_h}^3} (c_{a_h}c_{b_h} - \dot{c}_{a_h}\dot{c}_{b_h})(c_{a_h}c_{b_h} - \check{c}_{a_h}\check{c}_{b_h}),$$

where

$$\dot{c}_{a_h} \dot{c}_{b_h} = \frac{\gamma_h - \sqrt{\gamma_h^2 - 2\sigma^2(L+M)}}{L+M}, \quad (12.103)$$

$$\check{c}_{a_h} \check{c}_{b_h} = \frac{\gamma_h + \sqrt{\gamma_h^2 - 2\sigma^2(L+M)}}{L+M}. \quad (12.104)$$

Summarizing the preceding discussion, we have the following lemma:

Lemma 12.14 *If $\gamma_h \leq \underline{\gamma}^{\text{local-EMAP}}$, the EMAP free energy \hat{F}_h^{MAP} , defined by Eq. (12.82), is increasing for $c_{a_h} c_{b_h} > \varepsilon^2$, and therefore minimized at $c_{a_h} c_{b_h} = \varepsilon^2$. If $\gamma_h > \underline{\gamma}^{\text{local-EMAP}}$,*

$$\hat{F}_h^{\text{MAP}} \text{ is } \begin{cases} \text{increasing} & \text{for } \varepsilon^2 < c_{a_h} c_{b_h} < \dot{c}_{a_h} \dot{c}_{b_h}, \\ \text{decreasing} & \text{for } \dot{c}_{a_h} \dot{c}_{b_h} < c_{a_h} c_{b_h} < \check{c}_{a_h} \check{c}_{b_h}, \\ \text{increasing} & \text{for } \check{c}_{a_h} \check{c}_{b_h} < c_{a_h} c_{b_h} < +\infty, \end{cases}$$

and therefore has two (local) minima at $c_{a_h} c_{b_h} = \varepsilon^2$ and at $c_{a_h} c_{b_h} = \check{c}_{a_h} \check{c}_{b_h}$. Here $\dot{c}_{a_h} \dot{c}_{b_h}$ and $\check{c}_{a_h} \check{c}_{b_h}$ are defined by Eqs. (12.103) and (12.104), respectively.

When $\gamma_h > \underline{\gamma}^{\text{local-EMAP}}$, the EMAP free energy (12.82) at the local minima is

$$2\hat{F}_h^{\text{MAP}} = \begin{cases} 0 & \text{at } c_{a_h} c_{b_h} = \varepsilon^2, \\ (L+M)(\log \check{c}_{a_h} \check{c}_{b_h} - 1 + \chi) - \sigma^{-2} \left(\gamma_h - \frac{\sigma^2}{\check{c}_{a_h} \check{c}_{b_h}} \right)^2 & \text{at } c_{a_h} c_{b_h} = \check{c}_{a_h} \check{c}_{b_h}, \end{cases}$$

respectively. Since we assume that $\chi = -\log \varepsilon^2$ is an arbitrarily large constant ($\varepsilon^2 > 0$ is arbitrarily small), $c_{a_h} c_{b_h} = \varepsilon^2$ is always the global minimum.

Substituting Eq. (12.104) into Eq. (12.36) gives Eq. (12.86), which completes the proof of Theorem 12.12. \square

12.1.5 Proofs of Lemma 12.11 and Theorem 12.13

We first analyze the free energy for EPB-A learning. From Eq. (12.49), we have

$$c_{a_h} c_{b_h} = \frac{\sigma^2}{\sqrt{(\underline{\gamma}_h^{\text{PB-A}})^2 - M\sigma^2}}.$$

Therefore, if

$$\gamma_h \leq \sigma \sqrt{M},$$

there exists only the null solution (12.53) for any $c_{a_h}c_{b_h} > 0$, and therefore the free energy (12.81) is given by

$$\begin{aligned} 2\hat{F}_h^{\text{PB-A}} &= \min_{\hat{a}_h, \hat{b}_h \in \mathbb{R}, \hat{\sigma}_{a_h}^2 \geq \varepsilon^2, \hat{\sigma}_{b_h}^2 = \varepsilon^2} 2F_h \\ &= L \left(\log c_{b_h}^2 + \frac{\varepsilon^2}{c_{b_h}^2} - 1 + \chi \right). \end{aligned} \quad (12.105)$$

In the following, we consider the case where

$$\gamma_h > \sigma \sqrt{M}. \quad (12.106)$$

For $\varepsilon^2 \leq c_{a_h}c_{b_h} \leq \frac{\sigma^2}{\sqrt{\gamma_h^2 - M\sigma^2}}$, there still exists only the null solution (12.53) with the free energy given by Eq. (12.105). The positive solution (12.50) appears for $c_{a_h}c_{b_h} > \frac{\sigma^2}{\sqrt{\gamma_h^2 - M\sigma^2}}$ with the free energy given by

$$\begin{aligned} 2\hat{F}_h^{\text{PB-A}} &= \min_{\hat{a}_h, \hat{b}_h \in \mathbb{R}, \hat{\sigma}_{a_h}^2 \geq \varepsilon^2, \hat{\sigma}_{b_h}^2 = \varepsilon^2} 2F_h \\ &= \min_{\hat{a}_h, \hat{b}_h \in \mathbb{R}, \hat{\sigma}_{a_h}^2 \geq \varepsilon^2} \left\{ M \log \frac{c_{a_h}^2}{\hat{\sigma}_{a_h}^2} + L \log c_{b_h}^2 + \frac{\hat{b}_h^2}{c_{b_h}^2} \right. \\ &\quad \left. - \frac{2\hat{a}_h \hat{b}_h \gamma_h}{\sigma^2} + \left(\hat{a}_h^2 + M\hat{\sigma}_{a_h}^2 \right) \left(\frac{\hat{b}_h^2}{\sigma^2} + \frac{1}{c_{a_h}^2} \right) \right\} - (L + M) + L\chi \\ &= \min_{\hat{a}_h, \hat{b}_h \in \mathbb{R}, \hat{\sigma}_{a_h}^2 \geq \varepsilon^2} \left\{ M \log \frac{\hat{b}_h^2 + \sigma^2/c_{a_h}^2}{\sigma^2} + \frac{\hat{b}_h^2}{c_{b_h}^2} - \frac{\hat{b}_h^2 \gamma_h}{\hat{b}_h^2 + \sigma^2/c_{a_h}^2} \frac{2\gamma_h}{\sigma^2} + \frac{\hat{a}_h^2 + M\hat{\sigma}_{a_h}^2}{\hat{\sigma}_{a_h}^2} \right. \\ &\quad \left. + M \log c_{a_h}^2 + L \log c_{b_h}^2 - (L + M) + L\chi \right\} \\ &= \min_{\hat{b}_h \in \mathbb{R}} \left\{ M \log \left(\hat{b}_h^2 + \sigma^2/c_{a_h}^2 \right) + \frac{\hat{b}_h^2 + \sigma^2/c_{a_h}^2}{c_{b_h}^2} - \frac{\hat{b}_h^2 \gamma_h^2}{\sigma^2(\hat{b}_h^2 + \sigma^2/c_{a_h}^2)} \right\} \\ &\quad - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} + M \log c_{a_h}^2 + L \log c_{b_h}^2 - M \log \sigma^2 - L + L\chi \\ &= \min_{\hat{b}_h \in \mathbb{R}} \left\{ M \log \left(\hat{b}_h^2 + \sigma^2/c_{a_h}^2 \right) + \frac{\hat{b}_h^2 + \sigma^2/c_{a_h}^2}{c_{b_h}^2} + \frac{\gamma_h^2}{c_{a_h}^2 (\hat{b}_h^2 + \sigma^2/c_{a_h}^2)} \right\} \\ &\quad - \frac{\gamma_h^2}{\sigma^2} - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} + M \log c_{a_h}^2 + L \log c_{b_h}^2 - M \log \sigma^2 - L + L\chi. \end{aligned} \quad (12.107)$$

Here we used the conditions (12.61) and (12.62) for the PB-A solution. By substituting the other conditions (12.64) and (12.66) into Eq. (12.107), we have

$$\begin{aligned}
2\hat{F}_h^{\text{PB-A}} &= M \log \left(-M + \sqrt{M^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) + \frac{-M + \sqrt{M^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}}}{2} + \frac{M + \sqrt{M^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}}}{2} \\
&\quad - \frac{\gamma_h^2}{\sigma^2} - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} + M \log c_{a_h}^2 + (L + M) \log c_{b_h}^2 - M \log(2\sigma^2) - L + L\chi \\
&= M \log \left(-M + \sqrt{M^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) + \sqrt{M^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \\
&\quad - \frac{\gamma_h^2}{\sigma^2} + M \log c_{a_h}^2 + (L + M) \log c_{b_h}^2 - M \log(2\sigma^2) - L + L\chi.
\end{aligned} \tag{12.108}$$

The PB-B free energy can be derived in exactly the same way, and the result is symmetric to the PB-A free energy. Namely, if

$$\gamma_h \leq \sigma \sqrt{L},$$

there exists only the null solution (12.59) for any $c_{a_h} c_{b_h} > 0$ with the free energy given by

$$\begin{aligned}
2\hat{F}_h^{\text{PB-B}} &= \min_{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2 = \varepsilon^2, \widehat{\sigma}_{b_h}^2 \geq \varepsilon^2} 2F_h \\
&= M \left(\log c_{a_h}^2 + \frac{\varepsilon^2}{c_{a_h}^2} - 1 + \chi \right).
\end{aligned} \tag{12.109}$$

Assume that

$$\gamma_h > \sigma \sqrt{L}.$$

For $\varepsilon^2 \leq c_{a_h} c_{b_h} \leq \frac{\sigma^2}{\sqrt{\gamma_h^2 - L\sigma^2}}$, there still exists only the null solution (12.59) with the free energy given by Eq. (12.109). The positive solution (12.56) appears for $c_{a_h} c_{b_h} > \frac{\sigma^2}{\sqrt{\gamma_h^2 - L\sigma^2}}$ with the free energy given by

$$\begin{aligned}
2\hat{F}_h^{\text{PB-B}} &= \min_{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2 = \varepsilon^2, \widehat{\sigma}_{b_h}^2 \geq \varepsilon^2} 2F_h \\
&= L \log \left(-L + \sqrt{L^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) + \sqrt{L^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \\
&\quad - \frac{\gamma_h^2}{\sigma^2} + (L + M) \log c_{a_h}^2 + L \log c_{b_h}^2 - L \log(2\sigma^2) - M + M\chi.
\end{aligned} \tag{12.110}$$

By fixing the ratio between the prior covariances to $c_{a_h}/c_{b_h} = 1$ in Eqs. (12.105) and (12.108) through (12.110), and taking the posterior choice in Eq. (12.47) into account, we obtain Eqs. (12.83) and (12.84), which prove Lemma 12.11.

Let us minimize the free energy with respect to $c_{a_h} c_{b_h}$. When

$$\gamma_h \leq \sigma \sqrt{\max(L, M)},$$

the free energy is given by Eq. (12.84), and its derivative is given by

$$\frac{\partial 2\hat{F}_h^{\text{PB}}}{\partial (c_{a_h} c_{b_h})} = \frac{\min(L, M)}{c_{a_h}^2 c_{b_h}^2} (c_{a_h} c_{b_h} - \varepsilon^2).$$

This implies that the free energy \hat{F}_h^{PB} is increasing for $\varepsilon^2 < c_{a_h} c_{b_h} < \infty$, and therefore minimized at $c_{a_h} c_{b_h} = \varepsilon^2$.

Assume that

$$\gamma_h > \sigma \sqrt{\max(L, M)}.$$

In this case, the free energy is given by Eq. (12.83), which is continuous in the domain $\varepsilon^2 \leq c_{a_h} c_{b_h} < \infty$, and differentiable except at $c_{a_h} c_{b_h} = \frac{\sigma^2}{\sqrt{\gamma_h^2 - \max(L, M)\sigma^2}}$.

Although the continuity is not very obvious, one can verify it by checking the value at $c_{a_h} c_{b_h} = \frac{\sigma^2}{\sqrt{\gamma_h^2 - \max(L, M)\sigma^2}}$ for each case in Eq. (12.83). The continuity is also expected from the fact that the PB solution is continuous at the threshold $\gamma_h = \underline{\gamma}_h^{\text{PB}}$, i.e., the positive solution (Eq. (12.50) for PB-A and Eq. (12.56) for PB-B) converges to the null solution (Eq. (12.53) for PB-A and Eq. (12.59) for PB-B) when $\gamma_h \rightarrow \underline{\gamma}_h^{\text{PB-A}} + 0$.

The free energy (12.83) is the same as Eq. (12.84) for $\varepsilon^2 \leq c_{a_h} c_{b_h} \leq \frac{\sigma^2}{\sqrt{\gamma_h^2 - \max(L, M)\sigma^2}}$, and therefore increasing in $\varepsilon^2 < c_{a_h} c_{b_h} \leq \frac{\sigma^2}{\sqrt{\gamma_h^2 - \max(L, M)\sigma^2}}$. For $c_{a_h} c_{b_h} > \frac{\sigma^2}{\sqrt{\gamma_h^2 - \max(L, M)\sigma^2}}$, the derivative of the free energy with respect to $c_{a_h}^2 c_{b_h}^2$ is given by

$$\begin{aligned} \frac{\partial 2\hat{F}_h^{\text{PB}}}{\partial (c_{a_h}^2 c_{b_h}^2)} &= \frac{\min(L, M) + 2\max(L, M)}{2c_{a_h}^2 c_{b_h}^2} - \frac{4\gamma_h^2}{2c_{a_h}^4 c_{b_h}^4 \sqrt{\max(L, M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}}} + \frac{\sigma^2}{c_{a_h}^4 c_{b_h}^4} \\ &\quad - \frac{4 \max(L, M) \gamma_h^2}{2c_{a_h}^4 c_{b_h}^4 \sqrt{\max(L, M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \left(-\max(L, M) + \sqrt{\max(L, M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)} \\ &= \frac{\min(L, M) + 2\max(L, M)}{2c_{a_h}^2 c_{b_h}^2} + \frac{\sigma^2}{c_{a_h}^4 c_{b_h}^4} - \frac{4\gamma_h^2}{2c_{a_h}^4 c_{b_h}^4 \left(-\max(L, M) + \sqrt{\max(L, M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right)} \\ &= \frac{1}{2c_{a_h}^2 c_{b_h}^2} \left(L + M + 2 \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} - \sqrt{\max(L, M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) \\ &= \frac{1}{2c_{a_h}^4 c_{b_h}^4} \left((L + M)c_{a_h}^2 c_{b_h}^2 + 2\sigma^2 - c_{a_h} c_{b_h} \sqrt{\max(L, M)^2 c_{a_h}^2 c_{b_h}^2 + 4\gamma_h^2} \right), \end{aligned} \tag{12.111}$$

which has the same sign as

$$\begin{aligned}\tau(c_{a_h}^2 c_{b_h}^2) &= \left\{ (L + M) c_{a_h}^2 c_{b_h}^2 + 2\sigma^2 \right\}^2 - \left\{ c_{a_h} c_{b_h} \sqrt{\max(L, M)^2 c_{a_h}^2 c_{b_h}^2 + 4\gamma_h^2} \right\}^2 \\ &= (2LM + \min(L, M)^2) c_{a_h}^4 c_{b_h}^4 - 4(\gamma_h^2 - \sigma^2(L + M)) c_{a_h}^2 c_{b_h}^2 + 4\sigma^4.\end{aligned}\quad (12.112)$$

Eq. (12.112) is a quadratic function of $c_{a_h}^2 c_{b_h}^2$, being positive at $c_{a_h}^2 c_{b_h}^2 \rightarrow +0$ and at $c_{a_h}^2 c_{b_h}^2 \rightarrow +\infty$. The free energy has stationary points if and only if

$$\gamma_h \geq \sigma \sqrt{L + M + \sqrt{2LM + \min(L, M)^2}} \left(\equiv \underline{\gamma}^{\text{local-EPB}} \right), \quad (12.113)$$

because $\tau(c_{a_h}^2 c_{b_h}^2)$, which has the same sign as the derivative of the free energy, can be factorized (with real factors if and only if the condition (12.113) holds) as

$$\tau(c_{a_h}^2 c_{b_h}^2) = (2LM + \min(L, M)^2) (c_{a_h}^2 c_{b_h}^2 - \dot{c}_{a_h}^2 \dot{c}_{b_h}^2) (c_{a_h}^2 c_{b_h}^2 - \check{c}_{a_h}^2 \check{c}_{b_h}^2),$$

where

$$\dot{c}_{a_h}^2 \dot{c}_{b_h}^2 = 2 \cdot \frac{(\gamma_h^2 - \sigma^2(L+M)) - \sqrt{(\gamma_h^2 - \sigma^2(L+M))^2 - (2LM + \min(L, M)^2)\sigma^4}}{2LM + \min(L, M)^2}, \quad (12.114)$$

$$\check{c}_{a_h}^2 \check{c}_{b_h}^2 = 2 \cdot \frac{(\gamma_h^2 - \sigma^2(L+M)) + \sqrt{(\gamma_h^2 - \sigma^2(L+M))^2 - (2LM + \min(L, M)^2)\sigma^4}}{2LM + \min(L, M)^2}. \quad (12.115)$$

Summarizing the preceding discussion, we have the following lemma:

Lemma 12.15 *If $\gamma_h \leq \underline{\gamma}^{\text{local-EPB}}$, the EPB free energy \dot{F}_h^{PB} , defined by Eqs. (12.83) and (12.84), is increasing for $c_{a_h} c_{b_h} > \varepsilon^2$, and therefore minimized at $c_{a_h} c_{b_h} = \varepsilon^2$. If $\gamma_h > \underline{\gamma}^{\text{local-EPB}}$,*

$$\dot{F}_h^{\text{PB}} \text{ is } \begin{cases} \text{increasing} & \text{for } \varepsilon^2 < c_{a_h} c_{b_h} < \dot{c}_{a_h} \dot{c}_{b_h}, \\ \text{decreasing} & \text{for } \dot{c}_{a_h} \dot{c}_{b_h} < c_{a_h} c_{b_h} < \check{c}_{a_h} \check{c}_{b_h}, \\ \text{increasing} & \text{for } \check{c}_{a_h} \check{c}_{b_h} < c_{a_h} c_{b_h} < +\infty, \end{cases}$$

and therefore has two (local) minima at $c_{a_h} c_{b_h} = \varepsilon^2$ and at $c_{a_h} c_{b_h} = \check{c}_{a_h} \check{c}_{b_h}$. Here, $\dot{c}_{a_h} \dot{c}_{b_h}$ and $\check{c}_{a_h} \check{c}_{b_h}$ are defined by Eqs. (12.114) and (12.115), respectively.

When $\gamma_h > \underline{\gamma}^{\text{local-EPB}}$, the EPB free energy (12.83) at the null local solution $c_{a_h} c_{b_h} = \varepsilon^2$ is $2\dot{F}_h^{\text{PB}} = 0$, while the EPB free energy at the positive local solution $c_{a_h} c_{b_h} = \check{c}_{a_h} \check{c}_{b_h}$ contains the term $\min(L, M)\chi$ with $\chi = -\log \varepsilon^2$ assumed to be arbitrarily large. Therefore, the null solution is always the global minimum.

Substituting Eq. (12.115) into Eq. (12.46) gives Eq. (12.88), which completes the proof of Theorem 12.13. \square

12.1.6 Noise Variance Estimation

The noise variance σ^2 is unknown in many practical applications. In VB learning, minimizing the free energy (12.26) with respect also to σ^2 gives a reasonable estimator, with which perfect dimensionality recovery was proven in Chapter 8. Here, we investigate whether MAP learning and PB learning offer good noise variance estimators.

We first consider the nonempirical Bayesian variants where the prior covariances $\mathbf{C}_A, \mathbf{C}_B$ are treated as given constants. By using Lemma 12.10, we can write the MAP free energy with the variational parameters optimized, as a function of σ^2 , as follows:

$$\begin{aligned}
2\hat{F}^{\text{MAP}} &= LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^{\min(L,M)} \gamma_h^2}{\sigma^2} + \sum_{h=1}^H 2\hat{F}_h^{\text{MAP}} \\
&= LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^{\min(L,M)} \gamma_h^2}{\sigma^2} \\
&\quad + \sum_{h=1}^{\min(H,\bar{H})} \left\{ (L+M)(\log c_{a_h} c_{b_h} - 1 + \chi) - \sigma^{-2} \left(\gamma_h - \frac{\sigma^2}{c_{a_h} c_{b_h}} \right)^2 \right\} \\
&\quad + \sum_{h=\min(H,\bar{H})+1}^{\min(L,M)} (L+M) \left(\log c_{a_h} c_{b_h} + \frac{\varepsilon^2}{c_{a_h} c_{b_h}} - 1 + \chi \right) \\
&= LM \log \sigma^2 + \frac{\sum_{h=\min(H,\bar{H})+1}^{\min(L,M)} \gamma_h^2}{\sigma^2} + \sum_{h=1}^{\min(H,\bar{H})} \left(\frac{2\gamma_h}{c_{a_h} c_{b_h}} - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right) \\
&\quad + LM \log(2\pi) + \sum_{h=1}^{\min(L,M)} (L+M) \left(\log c_{a_h} c_{b_h} + \frac{\varepsilon^2}{c_{a_h} c_{b_h}} - 1 + \chi \right) \\
&= LM \log \sigma^2 + \frac{\sum_{h=\min(H,\bar{H})+1}^{\min(L,M)} \gamma_h^2}{\sigma^2} + \sum_{h=1}^{\min(H,\bar{H})} \left(\frac{2\gamma_h}{c_{a_h} c_{b_h}} - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} \right) + \text{const.}
\end{aligned} \tag{12.116}$$

for

$$\underline{\sigma}_{\bar{H}+1}^2 \leq \sigma^2 \leq \underline{\sigma}_{\bar{H}}^2, \tag{12.117}$$

where

$$\underline{\sigma}_h^2 = \begin{cases} \infty & \text{for } h = 0, \\ c_{a_h} c_{b_h} \gamma_h & \text{for } h = 1, \dots, \min(L, M), \\ 0 & \text{for } h = \min(L, M) + 1. \end{cases} \tag{12.118}$$

Assume that we use the full-rank model $H = \min(L, M)$, and expect the ARD property to find the correct rank. Under this setting, the free energy (12.116) can be arbitrarily small for $\sigma^2 \rightarrow +0$, because the first term diverges to $-\infty$, and the second term is equal to zero for $0 (= \underline{\sigma}_{\min(L,M)+1}^2) \leq \sigma^2 \leq c_{a_{\min(L,M)}} c_{b_{\min(L,M)}} \gamma_{\min(L,M)} (= \underline{\sigma}_{\min(L,M)}^2)$ (note that $\gamma_{\min(L,M)} > 0$ with probability 1). This leads to the following lemma:

Lemma 12.16 Assume that $H = \min(L, M)$ and C_A, C_B are given as constants. Then the MAP free energy with respect to σ^2 is (globally) minimized at

$$\widehat{\sigma}^{2 \text{ MAP}} \rightarrow +0.$$

The PB free energy behaves differently. By using Lemma 12.11, we can write the PB free energy with the variational parameters optimized, as a function of σ^2 , as follows:

$$\begin{aligned}
2\dot{F}^{\text{PB}} &= LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^{\min(L,M)} \gamma_h^2}{\sigma^2} + \sum_{h=1}^H 2\dot{F}_h^{\text{PB}} \\
&= LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^{\min(L,M)} \gamma_h^2}{\sigma^2} \\
&\quad + \sum_{h=1}^{\min(H,\bar{H})} \left\{ \frac{\min(L,M)+2\max(L,M)}{2} \log c_{a_h}^2 c_{b_h}^2 + \sqrt{\max(L,M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right. \\
&\quad - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} + \max(L,M) \log \left(-\max(L,M) + \sqrt{\max(L,M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) \\
&\quad \left. - \frac{\gamma_h^2}{\sigma^2} - \max(L,M) \log(2\sigma^2) + \min(L,M)(\chi - 1) \right\} \\
&\quad + \sum_{h=\min(H,\bar{H})+1}^{\min(L,M)} \min(L,M) \left(\log c_{a_h} c_{b_h} + \frac{\varepsilon^2}{c_{a_h} c_{b_h}} - 1 + \chi \right) \\
&= (\min(L,M) - \min(H,\bar{H})) \max(L,M) \log(2\sigma^2) + \frac{\sum_{h=\min(H,\bar{H})+1}^{\min(L,M)} \gamma_h^2}{\sigma^2} \\
&\quad + \sum_{h=1}^{\min(H,\bar{H})} \left\{ \max(L,M) \log c_{a_h}^2 c_{b_h}^2 + \sqrt{\max(L,M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right. \\
&\quad - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} + \max(L,M) \log \left(-\max(L,M) + \sqrt{\max(L,M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) \\
&\quad \left. + LM \log(\pi) + \sum_{h=1}^{\min(L,M)} \min(L,M) (\log c_{a_h} c_{b_h} - 1 + \chi) \right\} \\
&= (\min(L,M) - \min(H,\bar{H})) \max(L,M) \log(2\sigma^2) + \frac{\sum_{h=\min(H,\bar{H})+1}^{\min(L,M)} \gamma_h^2}{\sigma^2} \\
&\quad + \sum_{h=1}^{\min(H,\bar{H})} \left\{ \max(L,M) \log c_{a_h}^2 c_{b_h}^2 + \sqrt{\max(L,M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right. \\
&\quad - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} + \max(L,M) \log \left(-\max(L,M) + \sqrt{\max(L,M)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) \\
&\quad \left. + \text{const.} \right. \tag{12.119}
\end{aligned}$$

for

$$\underline{\sigma}_{\bar{H}+1}^{2 \text{ PB}} \leq \sigma^2 \leq \underline{\sigma}_{\bar{H}}^{2 \text{ PB}}, \tag{12.120}$$

where

$$\underline{\sigma}_h^2 \text{PB} = \begin{cases} \infty & \text{for } h = 0, \\ \frac{c_{a_h}^2 c_{b_h}^2}{2} \left(-\max(L, M) + \sqrt{\max(L, M)^2 + 4 \frac{\gamma_h}{c_{a_h}^2 c_{b_h}^2}} \right) & \text{for } h = 1, \dots, \min(L, M), \\ 0 & \text{for } h = \min(L, M) + 1. \end{cases} \quad (12.121)$$

We find a remarkable difference between the MAP free energy (12.116) and the PB free energy (12.119): unlike in the MAP free energy, the first log term in the PB free energy disappears for $0 = \underline{\sigma}_{\min(L,M)+1}^2 < \sigma^2 < \underline{\sigma}_{\min(L,M)}^2$, and therefore, the PB free energy does not diverge to $-\infty$ at $\sigma^2 \rightarrow +0$. We can actually prove that the noise variance estimator is lower-bounded by a positive value as follows. The PB free energy (12.121) is continuous, and, for $0 = \underline{\sigma}_{\min(L,M)+1}^2 < \sigma^2 < \underline{\sigma}_{\min(L,M)}^2$, it can be written as

$$2F^{\text{PB}} = -\frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} + \text{const.},$$

which is monotonically decreasing. This leads to the following lemma:

Lemma 12.17 *Assume that $H = \min(L, M)$ and C_A, C_B are given as constants. Then the noise variance estimator in PB learning is lower-bounded by*

$$\begin{aligned} \widehat{\sigma}^2 \text{MAP} &\geq \underline{\sigma}_{\min(L,M)}^2 \\ &= \frac{c_{a_{\min(L,M)}}^2 c_{b_{\min(L,M)}}^2}{2} \left(-\max(L, M) + \sqrt{\max(L, M)^2 + 4 \frac{\gamma_{\min(L,M)}}{c_{a_{\min(L,M)}}^2 c_{b_{\min(L,M)}}^2}} \right). \end{aligned} \quad (12.122)$$

We numerically investigated the behavior of the noise variance estimator by creating random observed matrices $\mathbf{V} = \mathbf{B}^* \mathbf{A}^{*\top} + \mathcal{E} \in \mathbb{R}^{L \times M}$, and depicting the VB, PB, and MAP free energies as functions of σ^2 with the variational parameters optimized. Figure 12.7 shows a typical case for $L = 20, M = 50, H^* = 2$ with the entries of $\mathbf{A}^* \in \mathbb{R}^{M \times H^*}$ and $\mathbf{B}^* \in \mathbb{R}^{L \times H^*}$ independently drawn from $\text{Gauss}_1(0, 1^2)$, and the entries of $\mathcal{E} \in \mathbb{R}^{L \times M}$ independently drawn from $\text{Gauss}_1(0, 0.3^2)$. We set the prior covariances to $c_{a_h} c_{b_h} = 1$. As Lemma 12.16 states, the global minimum of the MAP free energy is at $\sigma^2 \rightarrow +0$. Since no nontrivial local minimum is observed, local search gives the same trivial solution. On the other hand, the PB free energy has a minimum in the positive region $\sigma^2 > 0$ with probability 1, as Lemma 12.17 states. However, we empirically observed that PB learning tends to underestimate the noise

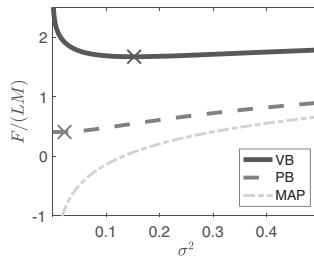


Figure 12.7 Free energy dependence on σ^2 . Crosses indicate nontrivial minima.

variance, as in Figure 12.7. Therefore, we cannot expect that the noise variance estimation works well in PB learning, either.

The situation is more complicated in the empirical Bayesian variants. Since the global EMAP estimator and the global EPB estimator, given any $\sigma^2 > 0$, are the null solution, the joint global optimization over all variational parameters and the hyperparameters results in $\sigma^2 = \sum_{h=1}^{\min(L,M)} \gamma_h^2 / (LM)$, regardless of observations—all observed signals are considered to be noise. If we adopt nontrivial local solutions as estimators, i.e., the local-EMAP estimator and the local-EPB estimator, the free energies are not continuous anymore as functions of σ^2 , because of the energy jump by the entropy factor χ of the pseudo-Dirac delta function. Even in that case, if we globally minimize the free energies with respect to σ^2 , the estimator contains no nontrivial local solution, because the null solutions cancel all entropy factors. As such, no reasonable way to estimate the noise variance has been found in EMAP learning and in EPB learning.

In the previous work on the TF model with PB learning (Chu and Ghahramani, 2009) and with EMAP learning (Mørup and Hansen, 2009), the noise variance was treated as a given constant. This was perhaps because the noise variance estimation failed, which is consistent with the preceding discussion.

12.2 More General Cases

Although extending the analysis for fully observed MF to more general cases is not easy in general, some basic properties can be shown. Specifically, this section shows that the global solutions for EMAP learning and EPB learning are also trivial and useless in the MF model with missing entries and in the TF model. Nevertheless, we experimentally show in Section 12.3 that local search for EMAP learning and EPB learning provides estimators that behave similarly to the EVB estimator.

12.2.1 Matrix Factorization with Missing Entries

The MF model with missing entries was introduced in Section 3.2. There, the likelihood (12.1) is replaced with

$$p(\mathbf{V}|\mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{1}{2\sigma^2} \left\| \mathcal{P}_\Lambda(\mathbf{V}) - \mathcal{P}_\Lambda(\mathbf{B}\mathbf{A}^\top) \right\|_{\text{Fro}}^2\right), \quad (12.123)$$

where Λ denotes the set of observed indices, and

$$(\mathcal{P}_\Lambda(\mathbf{V}))_{l,m} = \begin{cases} V_{l,m} & \text{if } (l, m) \in \Lambda, \\ 0 & \text{otherwise.} \end{cases}$$

The VB free energy is explicitly written as

$$\begin{aligned} 2F = & \#(\Lambda) \cdot \log(2\pi\sigma^2) + M \log \det(\mathbf{C}_A) + L \log \det(\mathbf{C}_B) \\ & - \sum_{m=1}^M \log \det(\widehat{\Sigma}_{A,m}) - \sum_{l=1}^L \log \det(\widehat{\Sigma}_{B,l}) - (L+M)H \\ & + \text{tr} \left\{ \mathbf{C}_A^{-1} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + \sum_{m=1}^M \widehat{\Sigma}_{A,m} \right) + \mathbf{C}_B^{-1} \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + \sum_{l=1}^L \widehat{\Sigma}_{B,l} \right) \right\} \\ & + \sigma^{-2} \sum_{(l,m) \in \Lambda} \left(V_{l,m} - 2V_{l,m} \widehat{\mathbf{a}}_m^\top \widehat{\mathbf{b}}_l + \text{tr} \left\{ \left(\widehat{\mathbf{a}}_m \widehat{\mathbf{a}}_m^\top + \widehat{\Sigma}_{A,m} \right) \left(\widehat{\mathbf{b}}_l \widehat{\mathbf{b}}_l^\top + \widehat{\Sigma}_{B,l} \right) \right\} \right), \end{aligned} \quad (12.124)$$

where $\#(\Lambda)$ denotes the number of observed entries.

We define the EMAP learning problem and the EPB learning problem by Eq. (12.77) with the free energy given by Eq. (12.124). The following holds:

Lemma 12.18 *The global solutions of EMAP learning and EPB learning for the MF model with missing entries, i.e., Eqs. (12.123), (12.2), and (12.3), are $\widehat{\mathbf{U}}^{\text{EMAP}} = \widehat{\mathbf{U}}^{\text{EPB}} = \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top = \mathbf{0}_{(L,M)}$, regardless of observations.*

Proof The posterior covariance for \mathbf{A} is clipped to $\widehat{\Sigma}_{A,m} = \varepsilon^2 \mathbf{I}_H$ in EPB-B learning, while the posterior covariance for \mathbf{B} is clipped to $\widehat{\Sigma}_{B,m} = \varepsilon^2 \mathbf{I}_H$ in EPB-A learning. In either case, one can make the second or the third term in the free energy (12.124) arbitrarily small to cancel the fourth or the fifth term by setting $\mathbf{C}_A = \varepsilon^2 \mathbf{I}_H$ or $\mathbf{C}_B = \varepsilon^2 \mathbf{I}_H$. Then, because of the terms in the third line of Eq. (12.124), which come from the prior distributions, it holds that $\widehat{\mathbf{A}} \rightarrow \mathbf{0}_{(M,H)}$ or $\widehat{\mathbf{B}} \rightarrow \mathbf{0}_{(M,H)}$ for $\varepsilon^2 \rightarrow +0$, which results in $\widehat{\mathbf{U}}^{\text{EPB}} = \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \rightarrow \mathbf{0}_{(L,M)}$. In EMAP learning, both posterior covariances are clipped to $\widehat{\Sigma}_{A,m} = \widehat{\Sigma}_{B,m} = \varepsilon^2 \mathbf{I}_H$. By the same argument as for EPB learning, we can show that $\widehat{\mathbf{U}}^{\text{EMAP}} = \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \rightarrow \mathbf{0}_{(L,M)}$, which completes the proof. \square

12.2.2 Tucker Factorization

The TF model was introduced in Section 3.3.1. The likelihood and the priors are given by

$$p(\mathcal{V}|\mathcal{G}, \{\mathbf{A}^{(n)}\}) \propto \exp\left(-\frac{\|\mathcal{V} - \mathcal{G} \times_1 \mathbf{A}^{(1)} \cdots \times_N \mathbf{A}^{(N)}\|^2}{2\sigma^2}\right), \quad (12.125)$$

$$p(\mathcal{G}) \propto \exp\left(-\frac{\text{vec}(\mathcal{G})^\top (\mathbf{C}_{G^{(N)}} \otimes \cdots \otimes \mathbf{C}_{G^{(1)}})^{-1} \text{vec}(\mathcal{G})}{2}\right), \quad (12.126)$$

$$p(\{\mathbf{A}^{(n)}\}) \propto \exp\left(-\frac{\sum_{n=1}^N \text{tr}(\mathbf{A}^{(n)} \mathbf{C}_{A^{(n)}}^{-1} \mathbf{A}^{(n)\top})}{2}\right), \quad (12.127)$$

where \otimes and $\text{vec}(\cdot)$ denote the *Kronecker product* and the *vectorization operator*, respectively. $\{\mathbf{C}_{G^{(n)}}\}$ and $\{\mathbf{C}_{A^{(n)}}\}$ are the prior covariances restricted to be diagonal, i.e.,

$$\begin{aligned} \mathbf{C}_{G^{(n)}} &= \mathbf{Diag}\left(c_{g_1^{(n)}}^2, \dots, c_{g_{H(n)}^{(n)}}^2\right), \\ \mathbf{C}_{A^{(n)}} &= \mathbf{Diag}\left(c_{a_1^{(n)}}^2, \dots, c_{a_{H(n)}^{(n)}}^2\right). \end{aligned}$$

We denote $\check{\mathbf{C}}_G = \mathbf{C}_{G^{(N)}} \otimes \cdots \otimes \mathbf{C}_{G^{(1)}}$.

The VB free energy is explicitly written as

$$\begin{aligned} 2F = & \left(\prod_{n=1}^N M^{(n)}\right) \log(2\pi\sigma^2) + \log \det(\check{\mathbf{C}}_G) + \sum_{n=1}^N M^{(n)} \log \det(\mathbf{C}_{A^{(n)}}) \\ & - \log \det(\widehat{\Sigma}_G) - \sum_{n=1}^N M^{(n)} \log \det(\widehat{\Sigma}_{A^{(n)}}) \\ & + \frac{\|\mathcal{V}\|^2}{\sigma^2} - \prod_{n=1}^N H^{(n)} - \prod_{n=1}^N (M^{(n)} H^{(n)}) \\ & + \text{tr}\left(\check{\mathbf{C}}_G^{-1} (\widetilde{\mathbf{g}} \widetilde{\mathbf{g}}^\top + \widehat{\Sigma}_G)\right) + \sum_{n=1}^N \text{tr}\left(\mathbf{C}_{A^{(n)}}^{-1} (\widehat{\mathbf{A}}^{(n)\top} \widehat{\mathbf{A}}^{(n)} + M^{(n)} \widehat{\Sigma}_{A^{(n)}})\right) \\ & - \frac{2}{\sigma^2} \check{\mathbf{v}}^\top (\widehat{\mathbf{A}}^{(N)} \otimes \cdots \otimes \widehat{\mathbf{A}}^{(1)}) \widetilde{\mathbf{g}} \\ & + \frac{1}{\sigma^2} \text{tr}\left\{ \left((\widehat{\mathbf{A}}^{(N)\top} \widehat{\mathbf{A}}^{(N)} + M^{(N)} \widehat{\Sigma}_{A^{(N)}}) \otimes \cdots \otimes (\widehat{\mathbf{A}}^{(1)\top} \widehat{\mathbf{A}}^{(1)} + M^{(1)} \widehat{\Sigma}_{A^{(1)}}) \right) \right. \\ & \quad \left. \cdot (\widetilde{\mathbf{g}} \widetilde{\mathbf{g}}^\top + \widehat{\Sigma}_G) \right\}. \end{aligned} \quad (12.128)$$

In the TF model, we refer as PB-G learning to the approximate Bayesian method where the posteriors for the factor matrices $\{\mathbf{A}^{(N)}\}$ are approximated

by the pseudo-Dirac delta function, and as PB-A learning to the one where the posterior for the core tensor \mathcal{G} is approximated by the pseudo-Dirac delta function. PB learning chooses the one giving a lower free energy from PB-G learning and PB-A learning. MAP learning approximates both posteriors by the pseudo-Dirac delta function. Note that the approach by Chu and Ghahramani (2009) corresponds to PB-G learning with the prior covariances fixed to $\mathbf{C}_{G^{(n)}} = \mathbf{C}_{A^{(n)}} = \mathbf{I}_{H^{(n)}}$ for $n = 1, \dots, N$, while the approach, called *ARD Tucker*, by Mørup and Hansen (2009) corresponds to EMAP learning with the prior covariances estimated from observations. In both approaches, the noise variance σ^2 was treated as a given constant.

Again the global solutions of EMAP learning and EPB learning are trivial and useless.

Lemma 12.19 *The global solutions of EMAP learning and EPB learning for the TF model, i.e., Eqs. (12.125) through (12.127), are $\widehat{\mathcal{U}}^{\text{EMAP}} = \widehat{\mathcal{U}}^{\text{EPB}} = \widehat{\mathcal{G}} \times_1 \widehat{\mathbf{A}}^{(1)} \cdots \times_N \widehat{\mathbf{A}}^{(N)} = \mathbf{0}_{(M^{(1)}, \dots, M^{(N)})}$, regardless of observations.*

Proof The posterior covariance for \mathcal{G} is clipped to $\widehat{\Sigma}_G = \varepsilon^2 \mathbf{I}_{\prod_{n=1}^N H^{(n)}}$ in EPB-A learning, while the posterior covariances for $\{\mathbf{A}^{(n)}\}$ are clipped to $\{\widehat{\Sigma}_{A^{(n)}} = \varepsilon^2 \mathbf{I}_{H^{(n)}}\}$ in EPB-G learning. In either case, one can make the second or the third term in the free energy (12.128) arbitrarily small to cancel the fourth or the fifth term by setting $\{\mathbf{C}_{G^{(n)}} = \varepsilon^2 \mathbf{I}_{H^{(n)}}\}$ or $\{\mathbf{C}_{A^{(n)}} = \varepsilon^2 \mathbf{I}_{H^{(n)}}\}$. Then, because of the terms in the fourth line of Eq. (12.128), which come from the prior distributions, it holds that $\widehat{\mathcal{G}} \rightarrow \mathbf{0}_{(H^{(1)}, \dots, H^{(N)})}$ or $\{\widehat{\mathbf{A}}^{(N)} \rightarrow \mathbf{0}_{(M^{(N)}, H^{(N)})}\}$ for $\varepsilon^2 \rightarrow +0$, which results in $\widehat{\mathcal{U}}^{\text{EPB}} = \mathbf{0}_{(M^{(1)}, \dots, M^{(N)})}$. In EMAP learning, both posterior covariances are clipped to $\widehat{\Sigma}_G = \varepsilon^2 \mathbf{I}_{\prod_{n=1}^N H^{(n)}}$ and $\{\widehat{\Sigma}_{A^{(n)}} = \varepsilon^2 \mathbf{I}_{H^{(n)}}\}$, respectively. By the same argument as for EPB learning, we can show that $\widehat{\mathcal{U}}^{\text{EMAP}} = \mathbf{0}_{(M^{(1)}, \dots, M^{(N)})}$, which completes the proof. \square

12.3 Experimental Results

In this section, we experimentally investigate the behavior of EMAP learning and EPB learning, in comparison with EVB learning. We start from the fully observed MF model, where we can assess how often local search finds the nontrivial local solution (derived in Section 12.1.3) rather than the global null solution. After that, we conduct experiments in collaborative filtering, where the MF model with missing entries is used, and in TF.

For local search, we adopted the standard iterative algorithm. The standard iterative algorithm for EVB learning has been derived in Chapter 3. The

standard iterative algorithms for EPB learning and EMAP learning, which can be derived simply by setting the derivatives of the corresponding free energies with respect to the unknown parameters to zero, similarly apply the stationary conditions in turn to update unknown parameters. For initialization, the entries of the mean parameters, e.g., $\widehat{\mathbf{A}}$, $\widehat{\mathbf{B}}$, and $\widehat{\mathcal{G}}$, were drawn from $\text{Gauss}_1(0, 1^2)$, while the covariance parameters were set to the identity, e.g., $\widehat{\Sigma}_G = \mathbf{I}_{\prod_{n=1}^N H^{(n)}}, \widehat{\Sigma}_{A^{(n)}} = \mathbf{C}_{G^{(n)}} = \mathbf{C}_{A^{(n)}} = \mathbf{I}_{H^{(n)}}$. We used this initialization scheme through all experiments in this section.

12.3.1 Fully Observed MF

We first conducted an experiment on an artificial (*ArtificialI*) data set, which was generated as follows. We randomly generated *true* matrices $\mathbf{A}^* \in \mathbb{R}^{M \times H^*}$ and $\mathbf{B}^* \in \mathbb{R}^{L \times H^*}$ such that each entry of \mathbf{A}^* and \mathbf{B}^* follows $\text{Gauss}_1(0, 1)$. An observed matrix $\mathbf{V} \in \mathbb{R}^{L \times M}$ was created by adding a noise subject to $\text{Gauss}_1(0, 1)$ to each entry of $\mathbf{B}^* \mathbf{A}^{*\top}$. Figures 12.8 through 12.10 show the free energy and the estimated rank over iterations in EVB learning, local-EPB

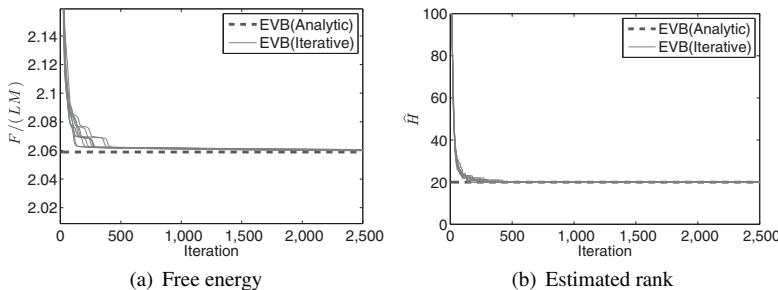


Figure 12.8 EVB learning on *ArtificialI* ($L = 100, M = 300, H^* = 20$).

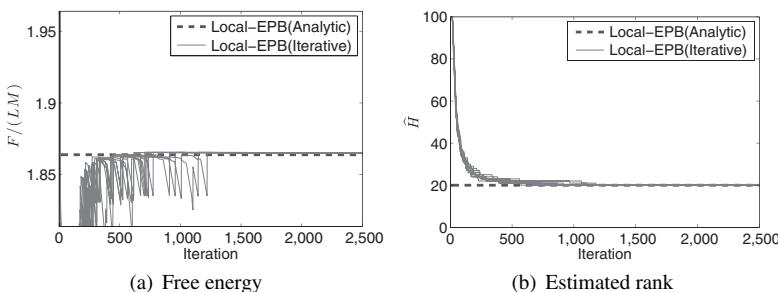
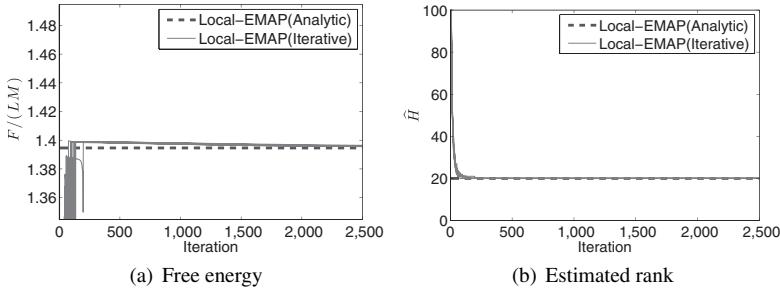


Figure 12.9 Local-EPB learning on *ArtificialI*.

Table 12.1 *Estimated rank in fully observed MF experiments.*

Data set	M	L	H^*	\widehat{H}^{EVB}		$\widehat{H}^{\text{local-EPB}}$		$\widehat{H}^{\text{local-EMAP}}$	
				Analytic	Iterative	Analytic	Iterative	Analytic	Iterative
<i>Artificial1</i>	300	100	20	20	20 (100%)	20	20 (100%)	20	20 (100%)
<i>Artificial2</i>	500	400	5	5	5 (100%)	8	8 (90%) 9 (10%)	5	5 (100%)
<i>Chart</i>	600	60	—	2	2 (100%)	2	2 (100%)	2	2 (100%)
<i>Glass</i>	214	9	—	1	1 (100%)	1	1 (100%)	1	1 (100%)
<i>Optical Digits</i>	5,620	64	—	10	10 (100%)	10	10 (100%)	6	6 (100%)
<i>Satellite</i>	6,435	36	—	2	2 (100%)	2	2 (100%)	1	1 (100%)

Figure 12.10 Local-EMAP learning on *Artificial1*.

learning, and local-EMAP learning, respectively, on the *Artificial1* data set with the data matrix size $L = 100$ and $M = 300$, and the true rank $H^* = 20$. The noise variance was assumed to be known, i.e., it was set to $\sigma^2 = 1$. We performed iterative local search 10 times, starting from different initial points, and each trial is plotted by a solid curve in the figures. The results computed by the analytic-form solutions for EVB learning (Theorem 6.13), local-EPB learning (Theorem 12.13), and local-EMAP learning (Theorem 12.12) were plotted as dashed lines. We can observe that iterative local search for EPB learning and EMAP learning tends to successfully find the nontrivial local solutions, although they are not global solutions.

We also conducted experiments on another artificial data set and benchmark data sets. The results are summarized in Table 12.1. *Artificial2* was created in the same way as *Artificial1*, but with $L = 400$, $M = 500$, and $H^* = 5$. The benchmark data sets were collected from the UCI repository (Asuncion and

Newman, 2007), on which we set the noise variance under the assumption that the signal to noise ratio is 0 db, following Mørup and Hansen (2009).

In the table, the estimated ranks by the analytic-form solution and by iterative local search are shown. The percentages for iterative local search indicate the frequencies over 10 trials. We observe the following: first, iterative local search tends to estimate the same rank as the analytic-form (local) solution; and second, the estimated rank tends to be consistent among EVB learning, local-EPB learning, and local-EMAP learning. Furthermore, on the artificial data sets, where the true rank is known, the rank is correctly estimated in most of the cases. Exceptions are *Artificial2*, where local-EPB learning overestimates the rank, and *Optical Digits* and *Satellite*, where local-EMAP learning estimates a smaller rank than the others. These phenomena can be explained by the theoretical implications in Section 12.1.3: in *Artificial2*, the ratio $\xi = H^*/\min(L, M) = 5/400$ between the true rank and the possible largest rank is small, which means that most of the singular components consist of noise. In such a case, local-EPB learning with its truncation threshold lower than MPUL tends to retain components purely consisting of noise (see Figure 12.6). In *Optical Digits* and *Satellite*, α ($= 64/5620$ for *Optical Digits* and $= 36/6435$ for *Satellite*) is extremely small, and therefore local-EMAP learning with its higher truncation threshold tends to discard more components than the others, as Figure 12.6 implies.

12.3.2 Collaborative Filtering

Next we conducted experiments in the collaborative filtering (CF) scenario, where the observed matrix has missing entries to be predicted by the MF model.

We generated an artificial (*ArtificialCF*) data set in the same way as the fully observed case for $L = 2,000, M = 5,000, H^* = 5$, and then masked 99% of the entries as missing values. We applied EVB learning, local-EPB learning, and local-EMAP learning to the MF model with missing entries, i.e., Eqs. (12.123), (12.2), and (12.3).⁴ Figure 12.11 shows the estimated rank and the *generalization error* over iterations for 10 trials, where the generalization error is defined as $GE = \|\mathcal{P}_{\Lambda'}(\mathbf{V}) - \mathcal{P}_{\Lambda'}(\widetilde{\mathbf{BA}}^\top)\|_{Fro}^2 / (\#(\Lambda') \sigma^2)$ for Λ' being the set of test indices.

⁴ Here we solve the EVB learning problem, the EPB learning problem, and the EMAP learning problem, respectively, by the standard iterative algorithms. However, we refer to the last two methods as local-EPB learning and local-EMAP learning, since we expect the local search algorithm to find not the global null solution but the nontrivial local solution.

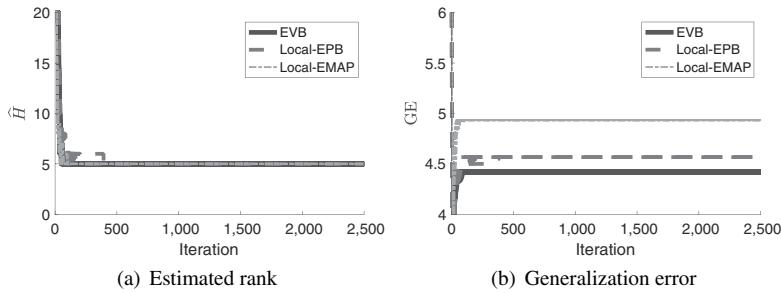


Figure 12.11 CF result on *ArtificialCF* ($L = 2,000$, $M = 5,000$, $H^* = 5$ with 99% missing ratio).

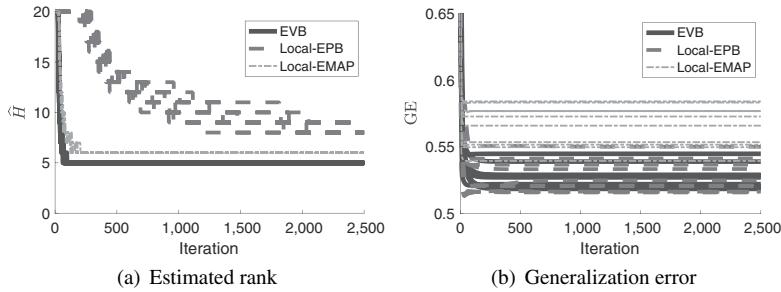


Figure 12.12 CF result on *MovieLens* ($L = 943$, $M = 1,682$ with 99% missing ratio).

We also conducted an experiment on the *MovieLens* data sets (with $L = 943$, $M = 1,682$).⁵ We randomly divided the observed entries into training entries and test entries, so that 99% of the entries are missing in the training phase. The test entries are used to evaluate the generalization error. Figure 12.12 shows the result in the same format as Figure 12.11.

We see that, on both data sets, local-EMAP learning tends to estimate a similar rank to EVB learning, while local-EPB learning tends to estimate a larger rank—a similar tendency to the fully observed case. In terms of the generalization error, local-EPB learning performs comparably to EVB learning, while local-EMAP learning performs slightly worse.

12.3.3 Tensor Factorization

Finally, we conducted experiments on TF. We created an artificial (*ArtificialTF*) data set, following Mørup and Hansen (2009): we drew a three-mode

⁵ www.grouplens.org/

Table 12.2 *Estimated rank (effective size of core tensor) in TF experiments.*

Data set	M	H^*	\widehat{H}^{EVB}	$\widehat{H}^{\text{local-EPB}}$	$\widehat{H}^{\text{local-EMAP}}$	$\widehat{H}^{\text{ARD-Tucker}}$
<i>ArtificialTF</i>	(30, 40, 50)	(3, 4, 5)	(3, 4, 5): 100%	(3, 4, 5): 100%	(3, 4, 5): 90% (3, 7, 5): 10%	(3, 4, 5): 100%
<i>FIA</i>	(12, 100, 89)	(3, 6, 4)	(3, 5, 3): 100%	(3, 5, 3): 100%	(3, 5, 2): 50% (4, 5, 2): 20% (5, 4, 2): 10% (4, 4, 2): 10% (8, 5, 2): 10%	(3, 4, 2): 70% (3, 5, 2): 10% (3, 7, 2): 10% (10, 4, 3): 10%

random tensor of the size $(M^{(1)}, M^{(1)}, M^{(1)}) = (30, 40, 50)$ with the signal components $(H^{(1)*}, H^{(2)*}, H^{(3)*}) = (3, 4, 5)$. The noise is added so that the signal-to-noise ratio is 0 db. We also used the *Flow Injection Analysis (FIA)* data set.⁶ Table 12.2 shows the estimated rank with frequencies over 10 trials. Here we also show the results by ARD Tucker with the ridge prior (Mørup and Hansen, 2009), performed with the code provided by the authors. Local-EMAP learning and ARD Tucker minimize exactly the same objective, and the slightly different results come from the differences in the local search algorithm (standard iterative vs. gradient descent) and in the initialization scheme.

We generally observe that all learning methods provide reasonable results, although local-EMAP learning, as well as ARD Tucker, is less stable than the others.

⁶ www.models.kvl.dk/datasets

Part IV

Asymptotic Theory

13

Asymptotic Learning Theory

Part IV is dedicated to asymptotic theory of variational Bayesian (VB) learning. In this part, “asymptotic limit” always means the limit when the number N of training samples goes to infinity. The main goal of asymptotic learning theory is to clarify the behavior of some statistics, e.g., the generalization error, the training error, and the Bayes free energy, which indicate how fast a learning machine can be trained as a function of the number of training samples and how the trained machine is biased to the training samples by overfitting. This provides the mathematical foundation of information criteria for model selection—a task to choose the degree of freedom of a statistical model based on observed training data. We can also evaluate the approximation accuracy of VB learning to full Bayesian learning in terms of the free energy, i.e., the gap between the VB free energy and the Bayes free energy, which corresponds to the tightness of the evidence lower-bound (ELBO) (see Section 2.1.1). In this first chapter of Part IV, we give an overview of asymptotic learning theory as the background for the subsequent chapters.

13.1 Statistical Learning Machines

A statistical learning machine consists of two fundamental components, a statistical model and a learning algorithm (Figure 13.1). The statistical model is denoted by a probabilistic distribution depending on some unknown parameters, and the learning algorithm estimates the unknown parameters from observed training samples. Before introducing asymptotic learning theory, we categorize statistical learning machines based on the model and the learning algorithm.

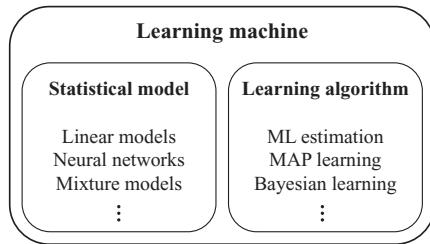


Figure 13.1 A statistical learning machine consists of a statistical model and a learning algorithm.

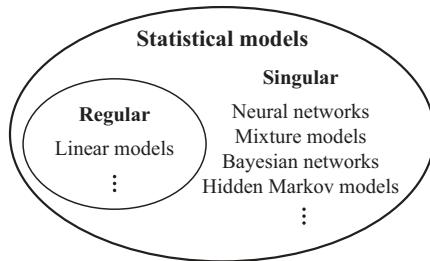


Figure 13.2 Statistical models are classified into regular models and singular models.

13.1.1 Statistical Models—Regular and Singular

We classify the statistical models into two classes, the *regular models* and the *singular models* (Figure 13.2). The regular models are identifiable (Definition 7.4 in Section 7.3.1), i.e.,

$$p(\mathbf{x}|\mathbf{w}_1) = p(\mathbf{x}|\mathbf{w}_2) \iff \mathbf{w}_1 = \mathbf{w}_2 \quad \text{for any } \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}, \quad (13.1)$$

and do not have *singularities* in the parameter space, i.e., the *Fisher information*

$$\mathbb{S}_+^D \ni \mathbf{F}(\mathbf{w}) = \int \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \left(\frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \right)^\top p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \quad (13.2)$$

is nonsingular (or full-rank) for any $\mathbf{w} \in \mathcal{W}$.

With a few additional assumptions, the regular models were analyzed under the *regularity conditions* (Section 13.4.1), which lead to the *asymptotic normality* of the distribution of the maximum likelihood (ML) estimator, and the asymptotic normality of the Bayes posterior distribution (Cramer,

1949; Sakamoto et al., 1986; van der Vaart, 1998). Based on those asymptotic normalities, a unified theory was established, clarifying the asymptotic behavior of generalization properties, which are common over all regular models, and over all reasonable learning algorithms, including ML learning, maximum a posteriori (MAP) learning, and Bayesian learning, as will be seen in Section 13.4.

On the other hand, analyzing singular models requires specific techniques for different models and different learning algorithms, and it was revealed that the asymptotic behavior of generalization properties depends on the model and the algorithm (Hartigan, 1985; Bickel and Chernoff, 1993; Takemura and Kuriki, 1997; Kuriki and Takemura, 2001; Amari et al., 2002; Hagiwara, 2002; Fukumizu, 2003; Watanabe, 2009). This is because the true parameter is at a singular point when the model size is larger than necessary to express the true distribution, and, in such cases, singularities affect the distribution of the ML estimator, as well as the Bayes posterior distribution even in the asymptotic limit. Consequently, the asymptotic normality, on which the regular learning theory relies, does not hold in singular models.

13.1.2 Learning Algorithms—Point Estimation and Bayesian Learning

When analyzing singular models, we also classify learning algorithms into two classes, point estimation and Bayesian learning (Figure 13.3). The point estimation methods, including ML learning and MAP learning, choose a single model (i.e., a single point in the parameter space) that maximizes a certain criterion such as the likelihood or the posterior probability, while Bayesian learning methods use an *ensemble* of models over the posterior distribution or its approximation.

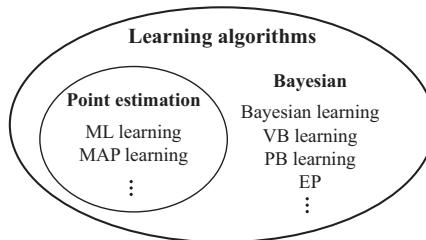


Figure 13.3 Learning algorithms are classified into point-estimation and Bayesian learning.

Unlike in the regular models, point estimation and Bayesian learning show different learning behavior in singular models. This is because how singularities affect the learning property depends on the learning methods. For example, as discussed in Chapter 7, strong nonuniformity of the density of the volume element leads to model-induced regularization (MIR) in Bayesian learning, while it does not affect point-estimation methods.

13.2 Basic Tools for Asymptotic Analysis

Here we introduce basic tools for asymptotic analysis.

13.2.1 Central Limit Theorem

Asymptotic learning theory heavily relies on the *central limit theorem*.

Theorem 13.1 (*Central limit theorem*) (van der Vaart, 1998) Let $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ be N i.i.d. samples from an arbitrary distribution with finite mean $\mu \in \mathbb{R}^D$ and finite covariance $\Sigma \in \mathbb{S}_{++}^D$, and let $\bar{\mathbf{x}} = N^{-1} \sum_{n=1}^N \mathbf{x}^{(n)}$ be their average. Then, the distribution of $\mathbf{z} = \sqrt{N}(\bar{\mathbf{x}} - \mu)$ converges to the Gaussian distribution with mean zero and covariance Σ ,¹ i.e.,

$$p(\mathbf{z}) \rightarrow \text{Gauss}_D(\mathbf{z}; \mathbf{0}, \Sigma) \quad \text{as} \quad N \rightarrow \infty. \quad (13.3)$$

Intuitively, Eq. (13.3) can be interpreted as

$$p(\bar{\mathbf{x}}) \rightarrow \text{Gauss}_D(\bar{\mathbf{x}}; \mu, N^{-1}\Sigma) \quad \text{as} \quad N \rightarrow \infty, \quad (13.4)$$

implying that the distribution of the average $\bar{\mathbf{x}}$ of i.i.d. random variables converges to the Gaussian distribution with mean μ and covariance $N^{-1}\Sigma$.

The central limit theorem implies the (weak) *law of large numbers*,² i.e., for any $\varepsilon > 0$,

$$\lim_{N \rightarrow \infty} \text{Prob}(\|\bar{\mathbf{x}} - \mu\| > \varepsilon) = 0. \quad (13.5)$$

13.2.2 Asymptotic Notation

We use the following *asymptotic notation*, a.k.a, *Bachmann–Landau notation*, to express the order of functions when the number N of samples goes to infinity:

¹ We consider weak topology in the space of distributions, i.e., $p(\mathbf{x})$ is identified with $r(\mathbf{x})$ if $\langle f(\mathbf{x}) \rangle_{p(\mathbf{x})} = \langle f(\mathbf{x}) \rangle_{r(\mathbf{x})}$ for any bounded continuous function $f(x)$. Convergence (of a random variable \mathbf{x}) in this sense is called *convergence in distribution*, *weak convergence*, or *convergence in law*, and denoted as $p(\mathbf{x}) \rightarrow r(\mathbf{x})$ or $\mathbf{x} \rightsquigarrow r(\mathbf{x})$ (van der Vaart, 1998).

² Convergence $\mathbf{x} \rightarrow \mu$ in the sense that $\lim_{N \rightarrow \infty} \text{Prob}(\|\mathbf{x} - \mu\| > \varepsilon) = 0, \forall \varepsilon > 0$ is called *convergence in probability*.

$O(f(N))$: A function such that $\limsup_{N \rightarrow \infty} |O(f(N))/f(N)| < \infty$,

$o(f(N))$: A function such that $\lim_{N \rightarrow \infty} o(f(N))/f(N) = 0$,

$\Omega(f(N))$: A function such that $\liminf_{N \rightarrow \infty} |\Omega(f(N))/f(N)| > 0$,

$\omega(f(N))$: A function such that $\lim_{N \rightarrow \infty} |\omega(f(N))/f(N)| = \infty$,

$\Theta(f(N))$: A function such that $\limsup_{N \rightarrow \infty} |\Theta(f(N))/f(N)| < \infty$

and $\liminf_{N \rightarrow \infty} |\Theta(f(N))/f(N)| > 0$.

Intuitively, as a function of N , $O(f(N))$ is a function of no greater order than $f(N)$, $o(f(N))$ is a function of less order than $f(N)$, $\Omega(f(N))$ is a function of no less order than $f(N)$, $\omega(f(N))$ is a function of greater order than $f(N)$, and $\Theta(f(N))$ is a function of the same order as $f(N)$. One thing we need to be careful of is that the upper-bounding notations, O and o , preserve after addition and subtraction, while lower-bounding notations, Ω and ω , as well as the both-sides-bounding notation Θ , do not necessarily preserve. For example, if $g_1(N) = \Theta(f(N))$ and $g_2(N) = \Theta(f(N))$ then $g_1(N) + g_2(N) = O(f(N))$, while it can happen that $g_1(N) + g_2(N) \neq \Theta(f(N))$ since the leading terms of $g_1(N)$ and $g_2(N)$ can coincide with each other with opposite signs and be canceled out.

For random variables, we use their probabilistic versions, O_p , o_p , Ω_p , ω_p , and Θ_p , for which the corresponding conditions hold in probability. For example, for i.i.d. samples $\{x^{(n)}\}_{n=1}^N$ from $\text{Gauss}_1(x; 0, 1^2)$, we can say that

$$\begin{aligned} x^{(n)} &= \Theta_p(1), \\ \bar{x} &= \frac{1}{N} \sum_{n=1}^N x^{(n)} = \Theta_p(N^{-1/2}), \\ \overline{x^2} &= \frac{1}{N} \sum_{n=1}^N (x^{(n)})^2 = 1 + \Theta_p(N^{-1/2}). \end{aligned}$$

Note that the second and the third equations are consequences from the central limit theorem (Theorem 13.1) applied to the samples $\{x^{(n)}\}$ that follow the Gaussian distribution, and to the samples $\{(x^{(n)})^2\}$ that follow the chi-squared distribution, respectively.

In this book, we express asymptotic approximation mostly by using asymptotic notation. To this end, we sometimes need to translate convergence of a random variable into an equation with asymptotic notation. Let x be a random variable depending on N , $r(x)$ be a distribution with finite mean and

covariance, and $f(\mathbf{x})$ be an arbitrary bounded continuous function. Then the following hold:

- If $p(\mathbf{x}) \rightarrow r(\mathbf{x})$, i.e., the distribution of \mathbf{x} converges to $r(\mathbf{x})$, then

$$\mathbf{x} = O_p(1) \quad \text{and} \quad \langle f(\mathbf{x}) \rangle_{p(\mathbf{x})} = \langle f(\mathbf{x}) \rangle_{r(\mathbf{x})} (1 + o(1)).$$

- If $\lim_{N \rightarrow \infty} \text{Prob}(\|\mathbf{x} - \mathbf{y}\| > \varepsilon) = 0$ for any $\varepsilon > 0$, then

$$\mathbf{x} = \mathbf{y} + o_p(1).$$

For example, the central limit theorem (13.3) implies that

$$\begin{aligned} \bar{\mathbf{x}} &= \boldsymbol{\mu} + O_p(N^{-1/2}), \\ \langle (\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \rangle_{p(\bar{\mathbf{x}})} &= N^{-1} \boldsymbol{\Sigma} + o(N^{-1}), \end{aligned}$$

while the law of large numbers (13.5) implies that

$$\bar{\mathbf{x}} = \boldsymbol{\mu} + o_p(1).$$

13.3 Target Quantities

Here we introduce target quantities to be analyzed in asymptotic learning theory.

13.3.1 Generalization Error and Training Error

Consider a statistical model $p(\mathbf{x}|\mathbf{w})$, where $\mathbf{x} \in \mathbb{R}^M$ is an observed random variable and $\mathbf{w} \in \mathbb{R}^D$ is a parameter to be estimated. Let $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})^\top \in \mathbb{R}^{N \times M}$ be N i.i.d. training samples taken from the true distribution $q(\mathbf{x})$. We assume *realizability*—the true distribution can be exactly expressed by the statistical model, i.e., $\exists \mathbf{w}^* \text{ s.t. } q(\mathbf{x}) = p(\mathbf{x}|\mathbf{w}^*)$, where \mathbf{w}^* is called the true parameter.

Learning algorithms estimate the parameter value \mathbf{w} or its posterior distribution given the training data $\mathcal{D} = \mathbf{X}$, and provide the predictive distribution $p(\mathbf{x}|X)$ for a new sample \mathbf{x} . For example, ML learning provides the predictive distribution given by

$$p^{\text{ML}}(\mathbf{x}|X) = p(\mathbf{x}|\widehat{\mathbf{w}}^{\text{ML}}), \tag{13.6}$$

where

$$\widehat{\mathbf{w}}^{\text{ML}} = \underset{\mathbf{w}}{\text{argmax}} \, p(X|\mathbf{w}) = \underset{\mathbf{w}}{\text{argmax}} \left(\prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) \right) \tag{13.7}$$

is the ML estimator, while Bayesian learning provides the predictive distribution given by

$$p^{\text{Bayes}}(\mathbf{x}|\mathbf{X}) = \langle p(\mathbf{x}|\mathbf{w}) \rangle_{p(\mathbf{w}|\mathbf{X})} = \int p(\mathbf{x}|\mathbf{w})p(\mathbf{w}|\mathbf{X})d\mathbf{w}, \quad (13.8)$$

where

$$p(\mathbf{w}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\mathbf{w})p(\mathbf{w})}{\int p(\mathbf{X}|\mathbf{w})p(\mathbf{w})d\mathbf{w}} \quad (13.9)$$

is the posterior distribution (see Section 1.1).

The *generalization error*, a criterion of generalization performance, is defined as the *Kullback–Leibler (KL) divergence* of the predictive distribution from the true distribution:

$$\text{GE}(\mathbf{X}) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{X})} d\mathbf{x}. \quad (13.10)$$

Its *empirical* variant,

$$\text{TE}(\mathbf{X}) = \frac{1}{N} \sum_{n=1}^N \log \frac{q(\mathbf{x}^{(n)})}{p(\mathbf{x}^{(n)}|\mathbf{X})}, \quad (13.11)$$

is called the *training error*, which is often used as an estimator of the generalization error. Note that, for the ML predictive distribution (13.6),

$$-N \cdot \text{TE}^{\text{ML}}(\mathbf{X}) = \sum_{n=1}^N \log \frac{p(\mathbf{x}^{(n)}|\widehat{\mathbf{w}}^{\text{ML}})}{q(\mathbf{x}^{(n)})} \quad (13.12)$$

corresponds to the *log-likelihood ratio*, an important statistic for statistical test, when the null hypothesis is true.

The generalization error (13.10) and the training error (13.11) are random variables that depend on realization of the training data \mathbf{X} . Taking the average over the distribution of training samples, we define deterministic quantities,

$$\overline{\text{GE}}(N) = \langle \text{GE}(\mathbf{X}) \rangle_{q(\mathbf{X})}, \quad (13.13)$$

$$\overline{\text{TE}}(N) = \langle \text{TE}(\mathbf{X}) \rangle_{q(\mathbf{X})}, \quad (13.14)$$

which are called the *average generalization error* and the *average training error*, respectively. Here $\langle \cdot \rangle_{q(\mathbf{X})}$ denotes the expectation value over the distribution of N training samples. The average generalization error and the average training error are scalar functions of the number N of samples, and represent generalization performance of a learning machine consisting of a statistical model and a learning algorithm. The optimality of Bayesian learning is proven in terms of the average generalization error (see Appendix D).

If a learning algorithm can successfully estimate the true parameter \mathbf{w}^* with reasonably small error, the average generalization error and the average training error converge to zero with the rate $\Theta(N^{-1})$ in the asymptotic limit.³ One of the main goals of asymptotic learning theory is to identify or bound the coefficients of their leading terms, i.e., λ and ν in the following asymptotic expansions:

$$\overline{\text{GE}}(N) = \lambda N^{-1} + o(N^{-1}), \quad (13.15)$$

$$\overline{\text{TE}}(N) = \nu N^{-1} + o(N^{-1}). \quad (13.16)$$

We call λ and ν the *generalization coefficient* and the *training coefficient*, respectively.

13.3.2 Bayes Free Energy

The marginal likelihood (defined by Eq. (1.6) in Chapter 1),

$$p(\mathbf{X}) = \int p(\mathbf{X}|\mathbf{w})p(\mathbf{w})d\mathbf{w} = \int p(\mathbf{w}) \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w})d\mathbf{w}, \quad (13.17)$$

is also an important quantity in Bayesian learning. As explained in Section 1.1.3, the marginal likelihood can be regarded as the likelihood of an *ensemble* of models—the set of model distributions with the parameters subject to the prior distribution. Following the concept of the “likelihood” in statistics, we can say that the ensemble of models giving the highest marginal likelihood is most likely. Therefore, we can perform model selection by maximizing the marginal likelihood (Efron and Morris, 1973; Schwarz, 1978; Akaike, 1980; MacKay, 1992; Watanabe, 2009). Maximizing the marginal likelihood (13.17) amounts to minimizing the *Bayes free energy*, defined by Eq. (1.60):

$$F^{\text{Bayes}}(\mathbf{X}) = -\log p(\mathbf{X}). \quad (13.18)$$

The Bayes free energy is a random variable depending on the training samples \mathbf{X} , and is of the order of $\Theta_p(N)$. However, the dominating part comes from the entropy of the true distribution, and does not depend on the statistical model nor the learning algorithm. In statistical learning theory, we therefore analyze the behavior of the *relative Bayes free energy*,

$$\widetilde{F}^{\text{Bayes}}(\mathbf{X}) = \log \frac{q(\mathbf{X})}{p(\mathbf{X})} = F^{\text{Bayes}}(\mathbf{X}) - NS_N(\mathbf{X}), \quad (13.19)$$

³ This holds if the estimator achieves a mean squared error in the same order as the *Cramér–Rao lower-bound*, i.e., $\langle \|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2 \rangle_{q(\mathbf{X})} \geq N^{-1} \text{tr}(\mathbf{F}^{-1}(\mathbf{w}^*))$, where \mathbf{F} is the Fisher information (13.2) at \mathbf{w}^* . The Cramér–Rao lower-bound holds for any unbiased estimator under the regularity conditions.

where

$$S_N(\mathbf{X}) = -\frac{1}{N} \sum_{n=1}^N \log q(\mathbf{x}^{(n)}) \quad (13.20)$$

is the *empirical entropy*. The negative of the relative Bayes free energy,

$$-\widetilde{F}^{\text{Bayes}}(\mathbf{X}) = \log \frac{\int p(\mathbf{w}) \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) d\mathbf{w}}{\prod_{n=1}^N q(\mathbf{x}^{(n)})}, \quad (13.21)$$

can be seen as an *ensemble* version of the log-likelihood ratio—the logarithm of the ratio between the *marginal likelihood* (alternative hypothesis) and the true likelihood (null hypothesis).

When the prior $p(\mathbf{w})$ is positive around the true parameter \mathbf{w}^* , the relative Bayes free energy (13.19) is known to be of the order of $\Theta(\log N)$ and can be asymptotically expanded as follows:

$$\widetilde{F}^{\text{Bayes}}(\mathbf{X}) = \lambda'^{\text{Bayes}} \log N + o_p(\log N), \quad (13.22)$$

where the coefficient of the leading term λ'^{Bayes} is called the *Bayes free energy coefficient*. Note that, although the relative Bayes free energy is a random variable depending on realization of the training data \mathbf{X} , the leading term in Eq. (13.22) is deterministic.

Let us define the average relative Bayes free energy over the distribution of training samples:

$$\overline{F}^{\text{Bayes}}(N) = \left\langle \widetilde{F}^{\text{Bayes}}(\mathbf{X}) \right\rangle_{q(\mathbf{X})} = \left\langle \log \frac{\prod_{n=1}^N q(\mathbf{x}^{(n)})}{\int p(\mathbf{w}) \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) d\mathbf{w}} \right\rangle_{q(\mathbf{X})}. \quad (13.23)$$

An interesting and useful relation can be found between the average Bayes generalization error and the average relative Bayes free energy (Levin et al., 1990):

$$\begin{aligned} \overline{\text{GE}}^{\text{Bayes}}(N) &= \left\langle \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p^{\text{Bayes}}(\mathbf{x}|\mathbf{X})} d\mathbf{x} \right\rangle_{q(\mathbf{X})} \\ &= \left\langle \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{\int p(\mathbf{x}|\mathbf{w}) p(\mathbf{w}|\mathbf{X}) d\mathbf{w}} d\mathbf{x} \right\rangle_{q(\mathbf{X})} \\ &= \left\langle \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{\int p(\mathbf{x}|\mathbf{w}) p(\mathbf{X}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w}} d\mathbf{x} \right\rangle_{q(\mathbf{X})} \\ &\quad - \left\langle \int q(\mathbf{x}) \log \frac{1}{\int p(\mathbf{X}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w}} d\mathbf{x} \right\rangle_{q(\mathbf{X})} \\ &= \left\langle \log \frac{q(\mathbf{x}) q(\mathbf{X})}{\int p(\mathbf{x}|\mathbf{w}) p(\mathbf{X}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w}} \right\rangle_{q(\mathbf{x}) q(\mathbf{X})} - \left\langle \log \frac{q(\mathbf{X})}{\int p(\mathbf{X}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w}} \right\rangle_{q(\mathbf{X})} \\ &= \overline{F}^{\text{Bayes}}(N+1) - \overline{F}^{\text{Bayes}}(N). \end{aligned} \quad (13.24)$$

The relation (13.24) combined with the asymptotic expansions, Eqs. (13.15) and (13.22), implies that the Bayes generalization coefficient and the Bayes free energy coefficient coincide with each other, i.e.,

$$\lambda'^{\text{Bayes}} = \lambda^{\text{Bayes}}. \quad (13.25)$$

Importantly, this relation holds for any statistical model, regardless of being regular or singular.

13.3.3 Target Quantities under Conditional Modeling

Many statistical models are for the regression or classification setting, where the model distribution $p(\mathbf{y}|\mathbf{x}, \mathbf{w})$ is the distribution of an output $\mathbf{y} \in \mathbb{R}^L$ conditional on an input $\mathbf{x} \in \mathbb{R}^M$ and an unknown parameter $\mathbf{w} \in \mathbb{R}^D$. The input is assumed to be given for all samples including the future test samples. Let $\mathcal{D} = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})\}$ be N i.i.d. training samples drawn from the true joint distribution $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}|\mathbf{x})q(\mathbf{x})$. As noted in Example 1.2 in Chapter 1, we can proceed with most computations without knowing the input distribution $q(\mathbf{x})$.

Let $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})^\top \in \mathbb{R}^{N \times M}$ and $\mathbf{Y} = (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)})^\top \in \mathbb{R}^{N \times L}$ separately summarize the inputs and the outputs in the training data. The predictive distribution, given as a conditional distribution on a new input \mathbf{x} as well as the whole training samples $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$, can usually be computed without any information on $q(\mathbf{X})$. For example, the ML predictive distribution is given as

$$p^{\text{ML}}(\mathbf{y}|\mathbf{x}, \mathcal{D}) = p(\mathbf{y}|\mathbf{x}, \widehat{\mathbf{w}}^{\text{ML}}), \quad (13.26)$$

where

$$\widehat{\mathbf{w}}^{\text{ML}} = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) \cdot q(\mathbf{X}) = \underset{\mathbf{w}}{\operatorname{argmax}} \left(\prod_{n=1}^N p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \mathbf{w}) \right) \quad (13.27)$$

is the ML estimator, while the Bayes predictive distribution is given as

$$p^{\text{Bayes}}(\mathbf{y}|\mathbf{x}, \mathcal{D}) = \langle p(\mathbf{y}|\mathbf{x}, \mathbf{w}) \rangle_{p(\mathbf{w}|\mathbf{X}, \mathbf{Y})} = \int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{Y}) d\mathbf{w}, \quad (13.28)$$

where

$$p(\mathbf{w}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}) \cdot q(\mathbf{X})}{\int p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w} \cdot q(\mathbf{X})} = \frac{p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w})}{\int p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}} \quad (13.29)$$

is the Bayes posterior distribution. Here (\mathbf{x}, \mathbf{y}) is a new input–output sample pair, assumed to be drawn from the true distribution $q(\mathbf{y}|\mathbf{x})q(\mathbf{x})$.

The generalization error (13.10), the training error (13.11), and the relative Bayes free energy (13.19) can be expressed as follows:

$$\text{GE}(\mathcal{D}) = \left\langle \log \frac{q(\mathbf{y}|\mathbf{x})q(\mathbf{x})}{p(\mathbf{y}|\mathbf{x}, \mathcal{D})q(\mathbf{x})} \right\rangle_{q(\mathbf{y}|\mathbf{x})q(\mathbf{x})} = \left\langle \log \frac{q(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}|\mathbf{x}, \mathcal{D})} \right\rangle_{q(\mathbf{y}|\mathbf{x})q(\mathbf{x})}, \quad (13.30)$$

$$\text{TE}(\mathcal{D}) = \frac{1}{N} \sum_{n=1}^N \log \frac{q(\mathbf{y}^{(n)}|\mathbf{x}^{(n)})q(\mathbf{x}^{(n)})}{p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \mathcal{D})q(\mathbf{x}^{(n)})} = \frac{1}{N} \sum_{n=1}^N \log \frac{q(\mathbf{y}^{(n)}|\mathbf{x}^{(n)})}{p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \mathcal{D})}, \quad (13.31)$$

$$\tilde{F}^{\text{Bayes}}(\mathcal{D}) = \log \frac{q(\mathbf{Y}|\mathbf{X})q(\mathbf{X})}{p(\mathbf{Y}|\mathbf{X})q(\mathbf{X})} = F^{\text{Bayes}}(\mathbf{Y}|\mathbf{X}) - NS_N(\mathbf{Y}|\mathbf{X}), \quad (13.32)$$

where

$$F^{\text{Bayes}}(\mathbf{Y}|\mathbf{X}) = \log \int p(\mathbf{w}) \prod_{n=1}^N p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \mathbf{w}) d\mathbf{w}, \quad (13.33)$$

$$S_N(\mathbf{Y}|\mathbf{X}) = -\frac{1}{N} \sum_{n=1}^N \log q(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}). \quad (13.34)$$

We can see that the input distribution $q(\mathbf{x})$ cancels out in most of the preceding equations, and therefore Eqs. (13.30) through (13.34) can be computed without considering $q(\mathbf{x})$. Note that in Eq. (13.30), $q(\mathbf{x})$ remains the distribution over which the expectation is taken. However, it is necessary only formally, and the expectation value does not depend on $q(\mathbf{x})$ (as long as the regularity conditions hold). The same applies to the average generalization error (13.13), the average training error (13.14), and the average relative Bayes free energy (13.23), where the expectation $\langle \cdot \rangle_{q(\mathbf{Y}|\mathbf{X})q(\mathbf{X})}$ over the distribution of the training samples is taken.

13.4 Asymptotic Learning Theory for Regular Models

In this section, we introduce the *regular learning theory*, which generally holds under the regularity conditions.

13.4.1 Regularity Conditions

The *regularity conditions* are defined for the statistical model $p(\mathbf{x}|\mathbf{w})$ parameterized by a finite-dimensional parameter vector $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^D$, and the true distribution $q(\mathbf{x})$. We include conditions for the prior distribution $p(\mathbf{w})$, which are necessary for analyzing MAP learning and Bayesian learning. There are variations, and we here introduce a (rough) simple set.

- (i) The statistical model $p(\mathbf{x}|\mathbf{w})$ is differentiable (as many times as necessary) with respect to the parameter $\mathbf{w} \in \mathcal{W}$ for any \mathbf{x} , and the differential operator and the integral operator are commutable.
- (ii) The statistical model $p(\mathbf{x}|\mathbf{w})$ is *identifiable*, i.e., Eq. (13.1) holds, and the Fisher information (13.2) is nonsingular (full-rank) at any $\mathbf{w} \in \mathcal{W}$.
- (iii) The support of $p(\mathbf{x}|\mathbf{w})$, i.e., $\{\mathbf{x} \in \mathcal{X}; p(\mathbf{x}|\mathbf{w}) > 0\}$, is common for all $\mathbf{w} \in \mathcal{W}$.
- (iv) The true distribution is *realizable* by the statistical model, i.e., $\exists \mathbf{w}^* \text{ s.t. } q(\mathbf{x}) = p(\mathbf{x}|\mathbf{w}^*)$, and the true parameter \mathbf{w}^* is an interior point of the domain \mathcal{W} .
- (v) The prior $p(\mathbf{w})$ is twice differentiable and bounded as $0 < p(\mathbf{w}) < \infty$ at any $\mathbf{w} \in \mathcal{W}$.

Note that the first three conditions are on the model distribution $p(\mathbf{x}|\mathbf{w})$, the fourth is on the true distribution $q(\mathbf{x})$, and the fifth is on the prior distribution $p(\mathbf{w})$.

An important consequence of the regularity conditions is that the log-likelihood can be Taylor-expanded about any $\bar{\mathbf{w}} \in \mathcal{W}$:

$$\begin{aligned} \log p(\mathbf{x}|\mathbf{w}) &= \log p(\mathbf{x}|\bar{\mathbf{w}}) + (\mathbf{w} - \bar{\mathbf{w}})^\top \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\bar{\mathbf{w}}} \\ &\quad + \frac{1}{2} (\mathbf{w} - \bar{\mathbf{w}})^\top \frac{\partial^2 \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \Big|_{\mathbf{w}=\bar{\mathbf{w}}} (\mathbf{w} - \bar{\mathbf{w}}) + O(\|\mathbf{w} - \bar{\mathbf{w}}\|^3). \end{aligned} \quad (13.35)$$

13.4.2 Consistency and Asymptotic Normality

We first show *consistency* and *asymptotic normality*, which hold in ML learning, MAP learning, and Bayesian learning.

Consistency of ML Estimator

The ML estimator is defined by

$$\hat{\mathbf{w}}^{\text{ML}} = \underset{\mathbf{w}}{\operatorname{argmax}} \log \left(\prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) \right) = \underset{\mathbf{w}}{\operatorname{argmax}} L_N(\mathbf{w}), \quad (13.36)$$

where

$$L_N(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}^{(n)}|\mathbf{w}). \quad (13.37)$$

By the law of large numbers (13.5), it holds that

$$L_N(\mathbf{w}) = L^*(\mathbf{w}) + o_p(1), \quad (13.38)$$

where

$$L^*(\mathbf{w}) = \langle \log p(\mathbf{x}|\mathbf{w}) \rangle_{p(\mathbf{x}|\mathbf{w}^*)}. \quad (13.39)$$

Identifiability of the statistical model guarantees that

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} L^*(\mathbf{w}) \quad (13.40)$$

is the unique maximizer. Eqs. (13.36), (13.38), and (13.40), imply the consistency of the ML estimator, i.e.,

$$\widehat{\mathbf{w}}^{\text{ML}} = \mathbf{w}^* + o_p(1). \quad (13.41)$$

Asymptotic Normality of the ML Estimator

Since the gradient $\partial L_N(\mathbf{w})/\partial \mathbf{w}$ is differentiable, the *mean value theorem*⁴ guarantees that there exists $\dot{\mathbf{w}} \in [\min(\widehat{\mathbf{w}}^{\text{ML}}, \mathbf{w}^*), \max(\widehat{\mathbf{w}}^{\text{ML}}, \mathbf{w}^*)]^D$ (where $\min(\cdot)$ and $\max(\cdot)$ operate elementwise) such that

$$\frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} = \frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*} + \frac{\partial^2 L_N(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \Big|_{\mathbf{w}=\dot{\mathbf{w}}} (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*). \quad (13.42)$$

By the definition (13.36) of the ML estimator and the differentiability of $L_N(\mathbf{w})$, the left-hand side of Eq. (13.42) is equal to zero, i.e.,

$$\frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} = \mathbf{0}. \quad (13.43)$$

The first term in the right-hand side of Eq. (13.42) can be written as

$$\frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*} = \frac{1}{N} \sum_{n=1}^N \frac{\partial \log p(\mathbf{x}^{(n)}|\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*}. \quad (13.44)$$

Since Eq. (13.40) and the differentiability of $L^*(\mathbf{w})$ imply that

$$\frac{\partial L^*(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*} = \left\langle \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} = \mathbf{0}, \quad (13.45)$$

the right-hand side of Eq. (13.44) is the average over N i.i.d. samples of the random variable

$$\frac{\partial \log p(\mathbf{x}^{(n)}|\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*},$$

⁴ The mean value theorem states that, for a differentiable function $f : [a, b] \mapsto \mathbb{R}$,

$$\exists c \in [a, b] \quad \text{s.t.} \quad \frac{df(x)}{dx} \Big|_{x=c} = \frac{f(b)-f(a)}{b-a}.$$

which follows a distribution with zero mean (Eq. (13.45)) and the covariance given by the Fisher information (13.2) at $\mathbf{w} = \mathbf{w}^*$, i.e.,

$$\mathbf{F}(\mathbf{w}^*) = \left\langle \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*} \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}}^\top \Big|_{\mathbf{w}=\mathbf{w}^*} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)}. \quad (13.46)$$

Therefore, according to the central limit theorem (Theorem 13.1), the distribution of the first term in the right-hand side of Eq. (13.42) converges to

$$P \left(\frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*} \right) \rightarrow \text{Gauss}_D \left(\frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*}; \mathbf{0}, N^{-1} \mathbf{F}(\mathbf{w}^*) \right). \quad (13.47)$$

The coefficient of the second term in the right-hand side of Eq. (13.42) satisfies

$$\frac{\partial^2 L_N(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \Big|_{\mathbf{w}=\hat{\mathbf{w}}} = \frac{\partial^2 L^*(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \Big|_{\mathbf{w}=\mathbf{w}^*} + o_p(1), \quad (13.48)$$

because of the law of large numbers and the consistency of the ML estimator, i.e., $[\min(\hat{\mathbf{w}}^{\text{ML}}, \mathbf{w}^*), \max(\hat{\mathbf{w}}^{\text{ML}}, \mathbf{w}^*)] \ni \hat{\mathbf{w}} \rightarrow \mathbf{w}^*$ since $\hat{\mathbf{w}}^{\text{ML}} \rightarrow \mathbf{w}^*$. Furthermore, the following relation holds under the regularity conditions (see Appendix B.2):

$$\frac{\partial^2 L^*(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \Big|_{\mathbf{w}=\mathbf{w}^*} = \left\langle \frac{\partial^2 \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \Big|_{\mathbf{w}=\mathbf{w}^*} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} = -\mathbf{F}(\mathbf{w}^*). \quad (13.49)$$

Substituting Eqs. (13.43), (13.48), and (13.49) into Eq. (13.42) gives

$$(\mathbf{F}(\mathbf{w}^*) + o_p(1))(\hat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*) = \frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*}. \quad (13.50)$$

Since the Fisher information is assumed to be invertible, Eq. (13.47) leads to the following theorem:

Theorem 13.2 (*Asymptotic normality of ML estimator*) Under the regularity conditions, the distribution of $\mathbf{v}^{\text{ML}} = \sqrt{N}(\hat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)$ converges to

$$p(\mathbf{v}^{\text{ML}}) \rightarrow \text{Gauss}_D(\mathbf{v}^{\text{ML}}; \mathbf{0}, \mathbf{F}^{-1}(\mathbf{w}^*)) \quad \text{as} \quad N \rightarrow \infty. \quad (13.51)$$

Theorem 13.2 implies that

$$\hat{\mathbf{w}}^{\text{ML}} = \mathbf{w}^* + O_p(N^{-1/2}). \quad (13.52)$$

13.4.3 Asymptotic Normality of the Bayes Posterior

The Bayes posterior can be written as follows:

$$p(\mathbf{w}|\mathbf{X}) = \frac{\exp(NL_N(\mathbf{w}) + \log p(\mathbf{w}))}{\int \exp(NL_N(\mathbf{w}) + \log p(\mathbf{w})) d\mathbf{w}}. \quad (13.53)$$

In the asymptotic limit, the factor $\exp(NL_N(\mathbf{w}))$ dominates the numerator, and the probability mass concentrates around the peak of $L_N(\mathbf{w})$ —the ML estimator $\widehat{\mathbf{w}}^{\text{ML}}$. The Taylor expansion of $L_N(\mathbf{w})$ about $\widehat{\mathbf{w}}^{\text{ML}}$ gives

$$\begin{aligned} L_N(\mathbf{w}) &\approx L_N(\widehat{\mathbf{w}}^{\text{ML}}) + (\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}})^\top \left. \frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \right|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} \\ &\quad + \frac{1}{2} (\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}})^\top \left. \frac{\partial^2 L_N(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \right|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} (\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}}) \\ &\approx L_N(\widehat{\mathbf{w}}^{\text{ML}}) - \frac{1}{2} (\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}})^\top \mathbf{F}(\mathbf{w}^*) (\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}}), \end{aligned} \quad (13.54)$$

where we used Eq. (13.43) and

$$\left. \frac{\partial^2 L_N(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \right|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} = -\mathbf{F}(\mathbf{w}^*) + o_p(1), \quad (13.55)$$

which is implied by the law of large numbers and the consistency of the ML estimator. Eqs. (13.53) and (13.54) imply that the Bayes posterior can be approximated by Gaussian in the asymptotic limit:

$$p(\mathbf{w}|X) \approx \text{Gauss}_D \left(\mathbf{w}; \widehat{\mathbf{w}}^{\text{ML}}, N^{-1} \mathbf{F}^{-1}(\mathbf{w}^*) \right).$$

The following theorem was derived with more accurate discussion.

Theorem 13.3 (*Asymptotic normality of the Bayes posterior*) (van der Vaart, 1998) *Under the regularity conditions, the (rescaled) Bayes posterior distribution $p(\mathbf{v}|X)$ where $\mathbf{v} = \sqrt{N}(\mathbf{w} - \mathbf{w}^*)$ converges to*

$$p(\mathbf{v}|X) \rightarrow \text{Gauss}_D \left(\mathbf{v}; \mathbf{v}^{\text{ML}}, \mathbf{F}^{-1}(\mathbf{w}^*) \right) \quad \text{as} \quad N \rightarrow \infty, \quad (13.56)$$

where $\mathbf{v}^{\text{ML}} = \sqrt{N}(\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)$.

Theorem 13.3 implies that

$$\widehat{\mathbf{w}}^{\text{MAP}} = \widehat{\mathbf{w}}^{\text{ML}} + o_p(N^{-1/2}), \quad (13.57)$$

$$\widehat{\mathbf{w}}^{\text{Bayes}} = \langle \mathbf{w} \rangle_{p(\mathbf{w}|X)} = \widehat{\mathbf{w}}^{\text{ML}} + o_p(N^{-1/2}), \quad (13.58)$$

which prove the consistency of the MAP estimator and the Bayesian estimator.

13.4.4 Generalization Properties

Now we analyze the generalization error and the training error in ML learning, MAP learning, and Bayesian learning, as well as the Bayes free energy. After that, we introduce information criteria for model selection, which were developed based on the asymptotic behavior of those quantities.

13.4.5 ML Learning

The generalization error of ML learning can be written as

$$\begin{aligned} \text{GE}_{\text{Regular}}^{\text{ML}}(X) &= \left\langle \log \frac{p(\mathbf{x}|\mathbf{w}^*)}{p(\mathbf{x}|\widehat{\mathbf{w}}^{\text{ML}})} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} \\ &= L^*(\mathbf{w}^*) - L^*(\widehat{\mathbf{w}}^{\text{ML}}) \end{aligned} \quad (13.59)$$

with $L^*(\mathbf{w})$ defined by Eq. (13.39). The Taylor expansion of the second term of Eq. (13.59) about the true parameter \mathbf{w}^* gives

$$\begin{aligned} L^*(\widehat{\mathbf{w}}^{\text{ML}}) &= L^*(\mathbf{w}^*) + (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)^\top \frac{\partial L^*(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*} \\ &\quad + \frac{1}{2} (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)^\top \frac{\partial^2 L^*(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \Big|_{\mathbf{w}=\mathbf{w}^*} (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*) + O(\|\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*\|^3) \\ &= L^*(\mathbf{w}^*) - \frac{1}{2} (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)^\top \mathbf{F}(\mathbf{w}^*) (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*) + O(\|\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*\|^3), \end{aligned} \quad (13.60)$$

where we used Eqs. (13.45) and (13.49) in the last equality. Substituting Eq. (13.60) into Eq. (13.59) gives

$$\text{GE}_{\text{Regular}}^{\text{ML}}(X) = \frac{1}{2} (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)^\top \mathbf{F}(\mathbf{w}^*) (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*) + O(\|\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*\|^3). \quad (13.61)$$

The asymptotic normality (Theorem 13.2) of the ML estimator implies that

$$\sqrt{N} \mathbf{F}^{\frac{1}{2}}(\mathbf{w}^*) (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*) \rightsquigarrow \text{Gauss}_D(\mathbf{0}, \mathbf{I}_D), \quad (13.62)$$

and that

$$O(\|\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*\|^3) = O_p(N^{-3/2}). \quad (13.63)$$

Eq. (13.62) implies that the distribution of $s = N(\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)^\top \mathbf{F}(\mathbf{w}^*) (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)$ converges to the *chi-squared distribution* with D degrees of freedom:⁵

$$p(s) \rightarrow \chi^2(s; D), \quad (13.64)$$

and therefore,

$$N \left\langle (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)^\top \mathbf{F}(\mathbf{w}^*) (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*) \right\rangle_{q(X)} = D + o(1). \quad (13.65)$$

Eqs. (13.61), (13.63), and (13.65) lead to the following theorem:

⁵ The chi-squared distribution with D degrees of freedom is the distribution of the sum of the squares of D i.i.d. samples drawn from $\text{Gauss}_1(0, 1^2)$. It is actually a special case of the Gamma distribution, and it holds that $\chi^2(x; D) = \text{Gamma}(x; D/2, 1/2)$. The mean and the variance are equal to D and $2D$, respectively.

Theorem 13.4 *The average generalization error of ML learning in the regular models can be asymptotically expanded as*

$$\overline{\text{GE}}_{\text{Regular}}^{\text{ML}}(N) = \left\langle \text{GE}_{\text{Regular}}^{\text{ML}}(X) \right\rangle_{q(X)} = \lambda_{\text{Regular}}^{\text{ML}} N^{-1} + o(N^{-1}), \quad (13.66)$$

where the generalization coefficient is given by

$$2\lambda_{\text{Regular}}^{\text{ML}} = D. \quad (13.67)$$

Interestingly, the leading term of the generalization error only depends on the parameter dimension or the degree of freedom of the statistical model.

The training error of ML learning can be analyzed in a similar fashion. It can be written as

$$\begin{aligned} \text{TE}_{\text{Regular}}^{\text{ML}}(X) &= N^{-1} \sum_{n=1}^N \log \frac{p(\mathbf{x}^{(n)}|\mathbf{w}^*)}{p(\mathbf{x}^{(n)}|\widehat{\mathbf{w}}^{\text{ML}})} \\ &= L_N(\mathbf{w}^*) - L_N(\widehat{\mathbf{w}}^{\text{ML}}) \end{aligned} \quad (13.68)$$

with $L_N(\mathbf{w})$ defined by Eq. (13.37). The Taylor expansion of the first term of Eq. (13.68) about the ML estimator $\widehat{\mathbf{w}}^{\text{ML}}$ gives

$$\begin{aligned} L_N(\mathbf{w}^*) &= L_N(\widehat{\mathbf{w}}^{\text{ML}}) + (\mathbf{w}^* - \widehat{\mathbf{w}}^{\text{ML}})^{\top} \frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} \\ &\quad + \frac{1}{2} (\mathbf{w}^* - \widehat{\mathbf{w}}^{\text{ML}})^{\top} \frac{\partial^2 L_N(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^{\top}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} (\mathbf{w}^* - \widehat{\mathbf{w}}^{\text{ML}}) + O(\|\mathbf{w}^* - \widehat{\mathbf{w}}^{\text{ML}}\|^3) \\ &= L_N(\widehat{\mathbf{w}}^{\text{ML}}) - \frac{1}{2} (\mathbf{w}^* - \widehat{\mathbf{w}}^{\text{ML}})^{\top} (\mathbf{F}(\mathbf{w}^*) + o_p(1)) (\mathbf{w}^* - \widehat{\mathbf{w}}^{\text{ML}}) \\ &\quad + O(\|\mathbf{w}^* - \widehat{\mathbf{w}}^{\text{ML}}\|^3), \end{aligned} \quad (13.69)$$

where we used Eqs. (13.43) and (13.55). Substituting Eq. (13.69) into Eq. (13.68) and applying Eq. (13.52), we have

$$\text{TE}_{\text{Regular}}^{\text{ML}}(X) = -\frac{1}{2} (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*)^{\top} \mathbf{F}(\mathbf{w}^*) (\widehat{\mathbf{w}}^{\text{ML}} - \mathbf{w}^*) + o_p(N^{-1}). \quad (13.70)$$

Thus, Eq. (13.70) together with Eq. (13.65) gives the following theorem:

Theorem 13.5 *The average training error of ML learning in the regular models can be asymptotically expanded as*

$$\overline{\text{TE}}_{\text{Regular}}^{\text{ML}}(N) = \left\langle \text{TE}_{\text{Regular}}^{\text{ML}}(X) \right\rangle_{q(X)} = \nu_{\text{Regular}}^{\text{ML}} N^{-1} + o(N^{-1}), \quad (13.71)$$

where the training coefficient is given by

$$2\nu_{\text{Regular}}^{\text{ML}} = -D. \quad (13.72)$$

Comparing Theorems 13.4 and 13.5, we see that the generalization coefficient and the training coefficient are antisymmetric with each other:

$$\lambda_{\text{Regular}}^{\text{ML}} = -\nu_{\text{Regular}}^{\text{ML}}.$$

13.4.6 MAP Learning

We first prove the following theorem:

Theorem 13.6 *For any (point-) estimator such that*

$$\widehat{\mathbf{w}} = \widehat{\mathbf{w}}^{\text{ML}} + o_p(N^{-1/2}), \quad (13.73)$$

it holds that

$$\text{GE}_{\text{Regular}}^{\widehat{\mathbf{w}}}(\mathbf{X}) = \left\langle \log \frac{p(\mathbf{x}|\mathbf{w}^*)}{p(\mathbf{x}|\widehat{\mathbf{w}})} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} = \text{GE}_{\text{Regular}}^{\text{ML}}(\mathbf{X}) + o_p(N^{-1}), \quad (13.74)$$

$$\text{TE}_{\text{Regular}}^{\widehat{\mathbf{w}}}(\mathbf{X}) = N^{-1} \sum_{n=1}^N \log \frac{p(\mathbf{x}^{(n)}|\mathbf{w}^*)}{p(\mathbf{x}^{(n)}|\widehat{\mathbf{w}})} = \text{TE}_{\text{Regular}}^{\text{ML}}(\mathbf{X}) + o_p(N^{-1}). \quad (13.75)$$

Proof The generalization error of the estimator $\widehat{\mathbf{w}}$ can be written as

$$\begin{aligned} \text{GE}_{\text{Regular}}^{\widehat{\mathbf{w}}}(\mathbf{X}) &= \left\langle \log \frac{p(\mathbf{x}|\mathbf{w}^*)}{p(\mathbf{x}|\widehat{\mathbf{w}})} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} \\ &= L^*(\mathbf{w}^*) - L^*(\widehat{\mathbf{w}}) \\ &= \text{GE}_{\text{Regular}}^{\text{ML}}(\mathbf{X}) + (L^*(\widehat{\mathbf{w}}^{\text{ML}}) - L^*(\widehat{\mathbf{w}})), \end{aligned} \quad (13.76)$$

where the second term can be expanded as

$$\begin{aligned} L^*(\widehat{\mathbf{w}}^{\text{ML}}) - L^*(\widehat{\mathbf{w}}) &= -(\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^{\text{ML}})^{\top} \frac{\partial L^*(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} \\ &\quad - \frac{1}{2} (\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^{\text{ML}})^{\top} \frac{\partial^2 L^*(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^{\top}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} (\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^{\text{ML}}) + O(\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^{\text{ML}}\|^3). \end{aligned} \quad (13.77)$$

Eqs. (13.45) and (13.52) (with the differentiability of $\partial L^*(\mathbf{w})/\partial \mathbf{w}$) imply that

$$\frac{\partial L^*(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} = \frac{\partial L^*(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}^*+O_p(N^{-1/2})} = O_p(N^{-1/2}),$$

with which Eqs. (13.73) and (13.77) lead to

$$L^*(\widehat{\mathbf{w}}^{\text{ML}}) - L^*(\widehat{\mathbf{w}}) = o_p(N^{-1}).$$

Substituting the preceding into Eq. (13.76) gives Eq. (13.74).

Similarly, the training error of the estimator $\widehat{\mathbf{w}}$ can be written as

$$\begin{aligned}\text{TE}_{\text{Regular}}^{\widehat{\mathbf{w}}}(\mathbf{X}) &= N^{-1} \sum_{n=1}^N \log \frac{p(\mathbf{x}^{(n)}|\mathbf{w}^*)}{p(\mathbf{x}^{(n)}|\widehat{\mathbf{w}})} \\ &= L_N(\mathbf{w}^*) - L_N(\widehat{\mathbf{w}}) \\ &= \text{TE}_{\text{Regular}}^{\text{ML}}(\mathbf{X}) + (L_N(\widehat{\mathbf{w}}^{\text{ML}}) - L_N(\widehat{\mathbf{w}})),\end{aligned}\quad (13.78)$$

where the second term can be expanded as

$$\begin{aligned}L_N(\widehat{\mathbf{w}}^{\text{ML}}) - L_N(\widehat{\mathbf{w}}) &= -(\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^{\text{ML}})^{\top} \left. \frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \right|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} \\ &\quad - \frac{1}{2} (\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^{\text{ML}})^{\top} \left. \frac{\partial^2 L_N(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^{\top}} \right|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} (\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^{\text{ML}}) + O(\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^{\text{ML}}\|^3).\end{aligned}\quad (13.79)$$

Eqs. (13.43), (13.73), and (13.79) lead to

$$L_N(\widehat{\mathbf{w}}^{\text{ML}}) - L_N(\widehat{\mathbf{w}}) = o_p(N^{-1}).$$

Substituting the preceding into Eq. (13.78) gives Eq. (13.75), which completes the proof. \square

Since the MAP estimator satisfies the condition (13.73) of Theorem 13.6 (see Eq. (13.57)), we obtain the following corollaries:

Corollary 13.7 *The average generalization error of MAP learning in the regular models can be asymptotically expanded as*

$$\overline{\text{GE}}_{\text{Regular}}^{\text{MAP}}(N) = \langle \text{GE}_{\text{Regular}}^{\text{MAP}}(\mathbf{X}) \rangle_{q(\mathbf{X})} = \lambda_{\text{Regular}}^{\text{MAP}} N^{-1} + o(N^{-1}), \quad (13.80)$$

where the generalization coefficient is given by

$$2\lambda_{\text{Regular}}^{\text{MAP}} = D. \quad (13.81)$$

Corollary 13.8 *The average training error of MAP learning in the regular models can be asymptotically expanded as*

$$\overline{\text{TE}}_{\text{Regular}}^{\text{MAP}}(N) = \langle \text{TE}_{\text{Regular}}^{\text{MAP}}(\mathbf{X}) \rangle_{q(\mathbf{X})} = \nu_{\text{Regular}}^{\text{MAP}} N^{-1} + o(N^{-1}), \quad (13.82)$$

where the training coefficient is given by

$$2\nu_{\text{Regular}}^{\text{MAP}} = -D. \quad (13.83)$$

13.4.7 Bayesian Learning

Eq. (13.58) and Theorem 13.6 imply that the Bayesian estimator also gives the same generalization and training coefficients as ML learning, if the *plug-in* predictive distribution $p(\mathbf{x}|\widehat{\mathbf{w}}^{\text{Bayes}})$, i.e., the model distribution with the Bayesian parameter plugged-in, is used for prediction. We can show that the proper Bayesian procedure with the predictive distribution $p(\mathbf{x}|X) = \langle p(\mathbf{x}|\mathbf{w}) \rangle_{p(\mathbf{w}|X)}$ also gives the same generalization and training coefficients.

We first prove the following theorem:

Theorem 13.9 *Let $r(\mathbf{w})$ be a (possibly approximate posterior) distribution of the parameter, of which the mean and the covariance satisfy the following:*

$$\widehat{\mathbf{w}} = \langle \mathbf{w} \rangle_{r(\mathbf{w})} = \mathbf{w}^* + O_p(N^{-1/2}), \quad (13.84)$$

$$\widehat{\Sigma}_{\mathbf{w}} = \left\langle (\mathbf{w} - \langle \mathbf{w} \rangle_{r(\mathbf{w})})(\mathbf{w} - \langle \mathbf{w} \rangle_{r(\mathbf{w})})^\top \right\rangle_{r(\mathbf{w})} = O_p(N^{-1}). \quad (13.85)$$

Then the generalization error and the training error of the predictive distribution $p(\mathbf{x}|X) = \langle p(\mathbf{x}|\mathbf{w}) \rangle_{r(\mathbf{w})}$ satisfy

$$\text{GE}_{\text{Regular}}^r(X) = \left\langle \log \frac{p(\mathbf{x}|\mathbf{w}^*)}{\langle p(\mathbf{x}|\mathbf{w}) \rangle_{r(\mathbf{w})}} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} = \text{GE}_{\text{Regular}}^{\widehat{\mathbf{w}}}(X) + o_p(N^{-1}), \quad (13.86)$$

$$\text{TE}_{\text{Regular}}^r(X) = N^{-1} \sum_{n=1}^N \log \frac{p(\mathbf{x}^{(n)}|\mathbf{w}^*)}{\langle p(\mathbf{x}^{(n)}|\mathbf{w}) \rangle_{r(\mathbf{w})}} = \text{TE}_{\text{Regular}}^{\widehat{\mathbf{w}}}(X) + o_p(N^{-1}), \quad (13.87)$$

where $\text{GE}_{\text{Regular}}^{\widehat{\mathbf{w}}}(X)$ and $\text{TE}_{\text{Regular}}^{\widehat{\mathbf{w}}}(X)$ are, respectively, the generalization error and the training error of the point estimator $\widehat{\mathbf{w}}$ (defined in Theorem 13.6).

Proof The predictive distribution can be expressed as

$$\begin{aligned} \langle p(\mathbf{x}|\mathbf{w}) \rangle_{r(\mathbf{w})} &= \langle \exp(\log p(\mathbf{x}|\mathbf{w})) \rangle_{r(\mathbf{w})} \\ &= \left\langle \exp \left(\log p(\mathbf{x}|\widehat{\mathbf{w}}) + (\mathbf{w} - \widehat{\mathbf{w}})^\top \frac{\partial \log p(\mathbf{x}|\mathbf{w}')}{\partial \mathbf{w}'} \Big|_{\mathbf{w}'=\widehat{\mathbf{w}}} \right. \right. \\ &\quad \left. \left. + \frac{1}{2} (\mathbf{w} - \widehat{\mathbf{w}})^\top \frac{\partial^2 \log p(\mathbf{x}|\mathbf{w}')}{\partial \mathbf{w}' \partial \mathbf{w}'^\top} \Big|_{\mathbf{w}'=\widehat{\mathbf{w}}} (\mathbf{w} - \widehat{\mathbf{w}}) + O(|\mathbf{w} - \widehat{\mathbf{w}}|^3) \right) \right\rangle_{r(\mathbf{w})} \\ &= p(\mathbf{x}|\widehat{\mathbf{w}}) \cdot \left\langle \left(1 + (\mathbf{w} - \widehat{\mathbf{w}})^\top \frac{\partial \log p(\mathbf{x}|\mathbf{w}')}{\partial \mathbf{w}'} \Big|_{\mathbf{w}'=\widehat{\mathbf{w}}} \right. \right. \\ &\quad \left. \left. + \frac{1}{2} (\mathbf{w} - \widehat{\mathbf{w}})^\top \frac{\partial \log p(\mathbf{x}|\mathbf{w}')}{\partial \mathbf{w}'} \Big|_{\mathbf{w}'=\widehat{\mathbf{w}}} \frac{\partial \log p(\mathbf{x}|\mathbf{w}')}{\partial \mathbf{w}'} \Big|_{\mathbf{w}'=\widehat{\mathbf{w}}}^\top (\mathbf{w} - \widehat{\mathbf{w}}) \right. \right. \\ &\quad \left. \left. + \frac{1}{2} (\mathbf{w} - \widehat{\mathbf{w}})^\top \frac{\partial^2 \log p(\mathbf{x}|\mathbf{w}')}{\partial \mathbf{w}' \partial \mathbf{w}'^\top} \Big|_{\mathbf{w}'=\widehat{\mathbf{w}}} (\mathbf{w} - \widehat{\mathbf{w}}) + O(|\mathbf{w} - \widehat{\mathbf{w}}|^3) \right) \right\rangle_{r(\mathbf{w})}. \end{aligned}$$

Here we first expanded $\log p(\mathbf{x}|\mathbf{w})$ about $\widehat{\mathbf{w}}$, and then expanded the exponential function (with $\exp(z) = 1 + z + z^2/2 + O(z^3)$).

Using the conditions (13.84) and (13.85) on $r(\mathbf{w})$, we have

$$\langle p(\mathbf{x}|\mathbf{w}) \rangle_{r(\mathbf{w})} = p(\mathbf{x}|\widehat{\mathbf{w}}) \cdot \left(1 + \frac{1}{2} \text{tr} \left(\widehat{\Sigma}_w \boldsymbol{\Phi}(\mathbf{x}; \widehat{\mathbf{w}}) \right) + O_p(N^{-3/2}) \right), \quad (13.88)$$

where

$$\boldsymbol{\Phi}(\mathbf{x}; \widehat{\mathbf{w}}) = \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}} \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}}^\top \Big|_{\mathbf{w}=\widehat{\mathbf{w}}} + \frac{\partial^2 \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}' \partial \mathbf{w}'^\top} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}}. \quad (13.89)$$

Therefore,

$$\begin{aligned} \left\langle \log \frac{\langle p(\mathbf{x}|\mathbf{w}) \rangle_{r(\mathbf{w})}}{p(\mathbf{x}|\widehat{\mathbf{w}})} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} &= \left\langle \log \left(1 + \frac{1}{2} \text{tr} \left(\widehat{\Sigma}_w \boldsymbol{\Phi}(\mathbf{x}; \widehat{\mathbf{w}}) \right) + O_p(N^{-3/2}) \right) \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} \\ &= \frac{1}{2} \left\langle \text{tr} \left(\widehat{\Sigma}_w \boldsymbol{\Phi}(\mathbf{x}; \widehat{\mathbf{w}}) \right) \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} + O_p(N^{-3/2}). \end{aligned} \quad (13.90)$$

Here we expanded the logarithm function (with $\log(1+z) = z + O(z^2)$), using the condition (13.85) on the covariance, i.e., $\widehat{\Sigma}_w = O_p(N^{-1})$.

The condition (13.84) on the mean, i.e., $\widehat{\mathbf{w}} = \mathbf{w}^* + O_p(N^{-1/2})$, implies that

$$\begin{aligned} \langle \boldsymbol{\Phi}(\mathbf{x}; \widehat{\mathbf{w}}) \rangle_{p(\mathbf{x}|\mathbf{w}^*)} &= \langle \boldsymbol{\Phi}(\mathbf{x}; \mathbf{w}^*) \rangle_{p(\mathbf{x}|\mathbf{w}^*)} + O_p(N^{-1/2}) \\ &= \mathbf{F}(\mathbf{w}^*) - \mathbf{F}(\mathbf{w}^*) + O_p(N^{-1/2}) \\ &= O_p(N^{-1/2}), \end{aligned} \quad (13.91)$$

where we used the definition of the Fisher information (13.46) and its equivalent expression (13.49) (under the regularity conditions). Eqs. (13.90) and (13.91) together with the condition (13.85) give

$$\left\langle \log \frac{\langle p(\mathbf{x}|\mathbf{w}) \rangle_{r(\mathbf{w})}}{p(\mathbf{x}|\widehat{\mathbf{w}})} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)} = O_p(N^{-3/2}),$$

which results in Eq. (13.86).

Similarly, by using the expression (13.88) of the predictive distribution, we have

$$\begin{aligned} N^{-1} \sum_{n=1}^N \log \frac{\langle p(\mathbf{x}^{(n)}|\mathbf{w}) \rangle_{r(\mathbf{w})}}{p(\mathbf{x}^{(n)}|\widehat{\mathbf{w}})} &= N^{-1} \sum_{n=1}^N \log \left(1 + \frac{1}{2} \text{tr} \left(\widehat{\Sigma}_w \boldsymbol{\Phi}(\mathbf{x}^{(n)}; \widehat{\mathbf{w}}) \right) + O_p(N^{-3/2}) \right) \\ &= \frac{1}{2} N^{-1} \sum_{n=1}^N \text{tr} \left(\widehat{\Sigma}_w \boldsymbol{\Phi}(\mathbf{x}^{(n)}; \widehat{\mathbf{w}}) \right) + O_p(N^{-3/2}). \end{aligned} \quad (13.92)$$

The law of large numbers (13.5) and Eq. (13.91) lead to

$$\begin{aligned} N^{-1} \sum_{n=1}^N \boldsymbol{\Phi}(\mathbf{x}^{(n)}; \widehat{\mathbf{w}}) &= \langle \boldsymbol{\Phi}(\mathbf{x}; \widehat{\mathbf{w}}) \rangle_{p(\mathbf{x}|\mathbf{w}^*)} + o_p(1) \\ &= o_p(1). \end{aligned}$$

Substituting the preceding and the condition (13.85) into Eq. (13.92) gives

$$N^{-1} \sum_{n=1}^N \log \frac{\langle p(\mathbf{x}^{(n)}|\mathbf{w}) \rangle_{r(\mathbf{w})}}{p(\mathbf{x}^{(n)}|\widehat{\mathbf{w}})} = o_p(N^{-1}),$$

which results in Eq. (13.87). This completes the proof. \square

The asymptotic normality of the Bayes posterior (Theorem 13.3), combined with the asymptotic normality of the ML estimator (Theorem 13.2), guarantees that the conditions (13.84) and (13.85) of Theorem 13.9 hold in Bayesian learning, which leads to the following corollaries:

Corollary 13.10 *The average generalization error of Bayesian learning in the regular models can be asymptotically expanded as*

$$\overline{\text{GE}}_{\text{Regular}}^{\text{Bayes}}(N) = \langle \text{GE}_{\text{Regular}}^{\text{Bayes}}(\mathbf{X}) \rangle_{q(\mathbf{X})} = \lambda_{\text{Regular}}^{\text{Bayes}} N^{-1} + o(N^{-1}), \quad (13.93)$$

where the generalization coefficient is given by

$$2\lambda_{\text{Regular}}^{\text{Bayes}} = D. \quad (13.94)$$

Corollary 13.11 *The average training error of Bayesian learning in the regular models can be asymptotically expanded as*

$$\overline{\text{TE}}_{\text{Regular}}^{\text{Bayes}}(N) = \langle \text{TE}_{\text{Regular}}^{\text{Bayes}}(\mathbf{X}) \rangle_{q(\mathbf{X})} = \nu_{\text{Regular}}^{\text{Bayes}} N^{-1} + o(N^{-1}), \quad (13.95)$$

where the training coefficient is given by

$$2\nu_{\text{Regular}}^{\text{Bayes}} = -D. \quad (13.96)$$

Asymptotic behavior of the Bayes free energy (13.18) was also analyzed (Schwarz, 1978; Watanabe, 2009). The Bayes free energy can be written as

$$\begin{aligned} F^{\text{Bayes}}(\mathbf{X}) &= -\log p(\mathbf{X}) \\ &= -\log \int p(\mathbf{w}) \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) d\mathbf{w} \\ &= -\log \int \exp(NL_N(\mathbf{w}) + \log p(\mathbf{w})) d\mathbf{w}, \end{aligned}$$

where the factor $\exp(NL_N(\mathbf{w}))$ dominates in the asymptotic limit. By using the Taylor expansion

$$L_N(\mathbf{w}) = L_N(\widehat{\mathbf{w}}^{\text{ML}}) + (\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}})^{\top} \frac{\partial L_N(\mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}}$$

$$\begin{aligned}
& + \frac{1}{2}(\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}})^{\top} \frac{\partial^2 L_N(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^{\top}} \Big|_{\mathbf{w}=\widehat{\mathbf{w}}^{\text{ML}}} (\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}}) + O\left(\|\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}}\|^3\right) \\
& = L_N(\widehat{\mathbf{w}}^{\text{ML}}) - \frac{1}{2}(\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}})^{\top} \left(F(\mathbf{w}^*) + o_p(1) \right) (\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}}) + O\left(\|\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}}\|^3\right),
\end{aligned}$$

we can approximate the Bayes free energy as follows:

$$\begin{aligned}
F^{\text{Bayes}}(\mathbf{X}) & \approx -NL_N(\widehat{\mathbf{w}}^{\text{ML}}) - \log \int \exp \left(-\frac{N}{2}(\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}})^{\top} F(\mathbf{w}^*) (\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}}) \right. \\
& \quad \left. + \log p(\mathbf{w}) \right) d\mathbf{w} \\
& = -NL_N(\widehat{\mathbf{w}}^{\text{ML}}) - \log \int \exp \left(-\frac{1}{2}\mathbf{v}^{\top} F(\mathbf{w}^*) \mathbf{v} \right. \\
& \quad \left. + \log p(\widehat{\mathbf{w}}^{\text{ML}} + N^{-1/2}\mathbf{v}) \right) \frac{d\mathbf{v}}{N^{D/2}} \\
& = -NL_N(\widehat{\mathbf{w}}^{\text{ML}}) + \frac{D}{2} \log N + O_p(1). \tag{13.97}
\end{aligned}$$

where $\mathbf{v} = \sqrt{N}(\mathbf{w} - \widehat{\mathbf{w}}^{\text{ML}})$ is a rescaled parameter, on which the integration was performed with $d\mathbf{v} = N^{D/2} d\mathbf{w}$.

Therefore, the relative Bayes free energy (13.19) can be written as

$$\begin{aligned}
\widetilde{F}^{\text{Bayes}}(\mathbf{X}) & = F^{\text{Bayes}}(\mathbf{X}) + NL_N(\mathbf{w}^*) \\
& \approx \frac{D}{2} \log N + N \left(L_N(\mathbf{w}^*) - L_N(\widehat{\mathbf{w}}^{\text{ML}}) \right) + O_p(1). \tag{13.98}
\end{aligned}$$

Here we used $S_N(\mathbf{X}) = -L_N(\mathbf{w}^*)$, which can be confirmed by their definitions (13.20) and (13.37). The second term in Eq. (13.98) is of the order of $O_p(1)$, because Eqs. (13.68), (13.70), and (13.52) imply that

$$L_N(\mathbf{w}^*) - L_N(\widehat{\mathbf{w}}^{\text{ML}}) = \text{TE}_{\text{Regular}}^{\text{ML}}(\mathbf{X}) = O_p(N^{-1}).$$

The following theorem was obtained with more rigorous discussion.

Theorem 13.12 (Watanabe, 2009) *The relative Bayes free energy for the regular models can be asymptotically expanded as*

$$\widetilde{F}_{\text{Regular}}^{\text{Bayes}}(\mathbf{X}) = F^{\text{Bayes}}(\mathbf{X}) - NS_N(\mathbf{X}) = \lambda'_{\text{Regular}}^{\text{Bayes}} \log N + O_p(1), \tag{13.99}$$

where the Bayes free energy coefficient is given by

$$2\lambda'_{\text{Regular}}^{\text{Bayes}} = D. \tag{13.100}$$

Note that Corollary 13.10 and Theorem 13.12 are consistent with Eq. (13.25), which holds for any statistical model.

13.4.8 Information Criteria

We have seen that the leading terms of the generalization error, the training error, and the relative Bayes free energy are proportional to the parameter dimension. Those results imply that how much a regular statistical model *overfits* training data mainly depends on the degrees of freedom of statistical models. Based on this insight, various *information criteria* were proposed for *model selection*.

Let us first recapitulate the model selection problem. Consider a $(D - 1)$ -degree polynomial regression model for one-dimensional input t and output y :

$$y = \sum_{d=1}^D w_d t^{d-1} + \varepsilon,$$

where ε denotes a noise. This model can be written as

$$y = \mathbf{w}^\top \mathbf{x} + \varepsilon,$$

where $\mathbf{w} \in \mathbb{R}^D$ is a parameter vector, and $\mathbf{x} = (1, t, t^2, \dots, t^{D-1})^\top$ is a transformed input vector. Suppose that the true distribution can be realized *just* with a $(D^* - 1)$ -degree polynomial:

$$y = \sum_{d=1}^{D^*} w_d^* t^{d-1} + \varepsilon = \mathbf{w}^{*\top} \mathbf{x}' + \varepsilon,$$

where $\mathbf{w}^* \in \mathbb{R}^{D^*}$ is the *true* parameter vector, and $\mathbf{x}' = (1, t, t^2, \dots, t^{D^*-1})^\top$.⁶

If we train a $(D - 1)$ -degree polynomial model for $D < D^*$, we expect poor generalization performance because the true distribution is not realizable, i.e., the model is too simple to express the true distribution. On the other hand, it was observed that if we train a model such that $D \gg D^*$, the generalization performance is also not optimal, because the unnecessarily high degree terms cause overfitting. Accordingly, finding an appropriate degree D of freedom, based on the observed data, is an important task, which is known as a model selection problem.

It would be a good strategy if we could choose D , which minimizes the generalization error (13.30). Ignoring the terms that do not depend on the model (or D), the generalization error can be written as

$$\text{GE}(\mathcal{D}) = - \int q(\mathbf{x})q(\mathbf{y}|\mathbf{x}) \log p(\mathbf{y}|\mathbf{x}, \mathcal{D}) d\mathbf{x} d\mathbf{y} + \text{const.} \quad (13.101)$$

⁶ By “just,” we mean that $w_{D^*}^* \neq 0$, and therefore the true distribution is not realizable with any $(D - 1)$ -degree polynomial for $D < D^*$.

Unfortunately, we cannot directly evaluate Eq. (13.101), since the true distribution $q(\mathbf{y}|\mathbf{x})$ is inaccessible. Instead, the training error (13.31),

$$\text{TE}(\mathcal{D}) = -\frac{1}{N} \sum_{n=1}^N \log p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \mathcal{D}) + \text{const.}, \quad (13.102)$$

is often used as an estimator for the generalization error. Although Eq. (13.102) is accessible, the training error is known to be a biased estimator for the generalization error (13.101). In fact, the training error does not reflect the negative effect of redundancy of the statistical model, and tends to be monotonically decreasing as the parameter dimension D increases.

Akaike's information criterion (AIC) (Akaike, 1974),

$$\text{AIC} = -2 \sum_{n=1}^N \log p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \hat{\mathbf{w}}^{\text{ML}}) + 2D, \quad (13.103)$$

was proposed as an estimator for the generalization error of ML learning with bias correction. Theorems 13.4 and 13.5 provide the bias between the generalization error and the training error as follows:

$$\begin{aligned} \langle \text{GE}_{\text{Regular}}^{\text{ML}}(\mathcal{D}) - \text{TE}_{\text{Regular}}^{\text{ML}}(\mathcal{D}) \rangle_{q(\mathcal{D})} &= \overline{\text{GE}}_{\text{Regular}}^{\text{ML}}(N) - \overline{\text{TE}}_{\text{Regular}}^{\text{ML}}(N) \\ &= \frac{\lambda_{\text{Regular}}^{\text{ML}} - v_{\text{Regular}}^{\text{ML}}}{N} + o(N^{-1}) \\ &= \frac{D}{N} + o(N^{-1}). \end{aligned} \quad (13.104)$$

Therefore, it holds that

$$\begin{aligned} \text{TE}_{\text{Regular}}^{\text{ML}}(\mathcal{D}) + \langle \text{GE}_{\text{Regular}}^{\text{ML}}(\mathcal{D}) - \text{TE}_{\text{Regular}}^{\text{ML}}(\mathcal{D}) \rangle_{q(\mathcal{D})} \\ &= \text{TE}_{\text{Regular}}^{\text{ML}}(\mathcal{D}) + \frac{D}{N} + o(N^{-1}) \\ &= \frac{\text{AIC}}{2N} - S_N(\mathbf{Y}|\mathbf{X}) + o(N^{-1}), \end{aligned} \quad (13.105)$$

where $S_N(\mathbf{Y}|\mathbf{X})$ is the (conditional) empirical entropy (13.34). Since the empirical entropy $S_N(\mathbf{Y}|\mathbf{X})$ does not depend on the model, Eq. (13.105) implies that minimizing AIC amounts to minimizing an asymptotically unbiased estimator for the generalization error.

Another strategy for model selection is to minimize an approximation to the Bayes free energy (13.33). Instead of performing integration for computing

the Bayes free energy, Schwarz (1978) proposed to minimize the *Bayesian information criterion (BIC)*:

$$\text{BIC} = \text{MDL} = -2 \sum_{n=1}^N \log p(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, \hat{\mathbf{w}}^{\text{ML}}) + D \log N. \quad (13.106)$$

Interestingly, an equivalent criterion, called the *minimum description length (MDL)* (Rissanen, 1986), was derived in the context of information theory in communication. The relation between BIC and the Bayes free energy can be directly found from the approximation (13.97), i.e., it holds that

$$\begin{aligned} F^{\text{Bayes}}(\mathbf{Y}|\mathbf{X}) &\approx -NL_N(\hat{\mathbf{w}}^{\text{ML}}) + \frac{D}{2} \log N + O_p(1) \\ &= \frac{\text{BIC}}{2} + O_p(1), \end{aligned} \quad (13.107)$$

and therefore minimizing BIC amounts to minimizing an approximation to the Bayes free energy.

The first terms of AIC (13.103) and BIC (13.106) are the *maximum log-likelihood*—the log-likelihood at the ML estimator—multiplied by -2 . The second terms, called penalty terms, penalize high model complexity, which explicitly work as *Occam's razor* (MacKay, 1992) to prune off irrelevant degrees of freedom of the statistical model. AIC, BIC, and MDL are easily computable and have shown their usefulness in many applications. However, their derivations rely on the fact that the generalization coefficient, the training coefficient, and the free energy coefficient depend only on the parameter dimension *under the regularity conditions*. Actually, it has been revealed that, in singular models, those coefficients depend not only on the parameter dimension D but also on the true distribution.

13.5 Asymptotic Learning Theory for Singular Models

Many popular statistical models do not satisfy the regularity conditions. For example, neural networks, matrix factorization, mixture models, hidden Markov models, and Bayesian networks are all unidentifiable and have singularities, where the Fisher information is singular, in the parameter space.

As discussed in Chapter 7, the true parameter is on a singular point when the true distribution is realizable with a model with parameter dimension smaller than the used model, i.e., when the model has redundant components for expressing the true distribution. In such cases, the likelihood cannot be Taylor-expanded about the true parameter, and the asymptotic normality does not hold.

Consequently, the regular learning theory, described in Section 13.4, cannot be applied to singular models.

In this section, we first give intuition on how singularities affect generalization properties, and then introduce asymptotic theoretical results on ML learning and Bayesian learning. After that, we give an overview of asymptotic theory of VB learning, which will be described in detail in the subsequent chapters.

13.5.1 Effect of Singularities

Two types of effects of singularities have been observed, which will be detailed in the following subsections.

Basis Selection Effect

Consider a regression model for one-dimensional input $x \in [-10, 10]$ and output $y \in \mathbb{R}$ with H radial basis function (RBF) units:

$$p(y|x, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y - f(x; \mathbf{a}, \mathbf{b}, \mathbf{c}))^2\right), \quad (13.108)$$

$$\text{where } f(x; \mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_{h=1}^H \rho_h(x; a_h, b_h, c_h^2). \quad (13.109)$$

Each RBF unit in Eq. (13.109) is a weighted Gaussian RBF,

$$\rho_h(x; a_h, b_h, c_h^2) = a_h \cdot \text{Gauss}_1(x; b_h, c_h^2) = \frac{a_h}{\sqrt{2\pi c_h^2}} \exp\left(-\frac{(x - b_h)^2}{2c_h^2}\right),$$

controlled by a weight parameter $a_h \in \mathbb{R}$, a mean parameter $b_h \in \mathbb{R}$, and a scale parameter $c_h^2 \in \mathbb{R}_{++}$. Treating the noise variance σ^2 in Eq. (13.108) as a known constant, the parameters to be estimated are summarized as $\mathbf{w} = (\mathbf{a}^\top, \mathbf{b}^\top, \mathbf{c}^\top)^\top \in \mathbb{R}^{3H}$, where $\mathbf{a} = (a_1, \dots, a_H)^\top \in \mathbb{R}^H$, $\mathbf{b} = (b_1, \dots, b_H)^\top \in \mathbb{R}^H$, and $\mathbf{c} = (c_1^2, \dots, c_H^2)^\top \in \mathbb{R}_{++}^H$. Figure 13.4(a) shows an example of the RBF regression function (13.109) for $H = 2$.

Apparently, the model (13.108) is unidentifiable, and has singularities—since $\rho_h(x; 0, b_h, c_h^2) = 0$ for any $b_h \in \mathbb{R}$, $c_h^2 \in \mathbb{R}_{++}$, the (b_h, c_h^2) half-space at $a_h = 0$ is an unidentifiable set, on which the Fisher information is singular (see Figure 13.5).⁷ Accordingly, we call the model (13.108) a singular RBF regression model, of which the parameter dimension is equal to $D_{\text{sin-RBF}} = 3H$.

⁷ More unidentifiable sets can exist, depending on the other RBF units. See Section 7.3.1 for details on identifiability.

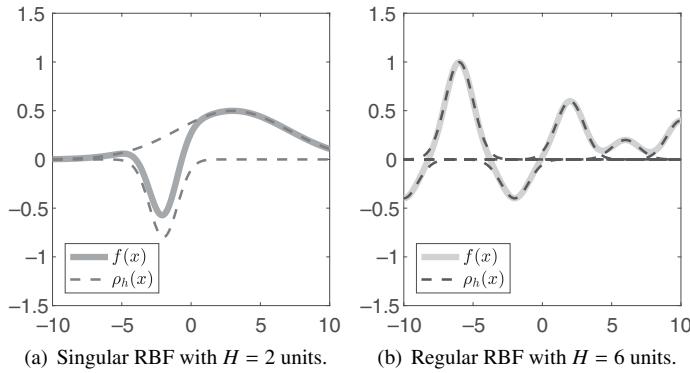


Figure 13.4 Examples (solid curves) of the singular RBF regression function (13.109) and the regular RBF regression function (13.111). Each RBF unit $\rho_h(x)$ is depicted as a dashed curve.

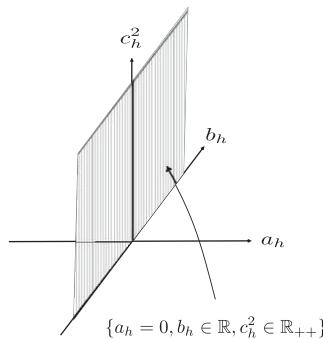


Figure 13.5 Singularities of the RBF regression model (13.108).

Let us consider another RBF regression model

$$p(y|x, \mathbf{a}) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(y - f(x; \mathbf{a}))^2\right), \quad (13.110)$$

$$\text{where } f(x; \mathbf{a}) = \sum_{h=1}^H \check{\rho}_h(x; a_h) = \sum_{h=1}^H \rho_h(x; a_h, \check{b}_h, \check{c}_h^2). \quad (13.111)$$

Unlike the singular RBF model (13.108), we here treat the mean parameters $\check{\mathbf{b}} = (\check{b}_1, \dots, \check{b}_H)^\top$ and the scale parameters $\check{\mathbf{c}} = (\check{c}_1^2, \dots, \check{c}_H^2)^\top$ as fixed constants, and only estimate the weight parameters $\mathbf{a} = (a_1, \dots, a_H)^\top \in \mathbb{R}^H$.

Let us set the means and the scales as follows, so that the model covers the input domain $[-10, 10]$:

$$\check{b}_h = -10 + 20 \cdot \frac{h-1}{H-1}, \quad (13.112)$$

$$\check{c}_h^2 = 1. \quad (13.113)$$

Figure 13.4(b) shows an example of the RBF regression function (13.111) for $H = 6$. Clearly, it holds that $\check{\rho}_h(x; a_h) \neq \check{\rho}_h(x; a'_h)$ if $a_h \neq a'_h$, and therefore the model is identifiable. The other regularity conditions (summarized in Section 13.4.1) on the model distribution $p(y|x, \mathbf{a})$ are also satisfied. Accordingly, we call the model (13.110) a regular RBF regression model, of which the parameter dimension is equal to $D_{\text{reg-RBF}} = H$.

Now we investigate difference in learning behavior between the singular RBF model (13.108) and the regular RBF model (13.110). Figure 13.6 shows trained regression functions (by ML learning) from $N = 50$ samples (shown as crosses) generated from the regression model,

$$q(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y - f^*(x))^2\right), \quad (13.114)$$

with the following true functions:

- (i) *poly*: Polynomial function $f^*(x) = -0.002x^3$.
- (ii) *cos*: Cosine function $f^*(x) = \cos(0.5x)$.
- (iii) *tanh*: Tangent hyperbolic function $f^*(x) = \tanh(-0.5x)$.
- (iv) *sin-sig*: Sine times sigmoid function $f^*(x) = \sin(x) \cdot \frac{1}{1+e^{-x}}$.
- (v) *sin-alg*: Sine function aligned for the regular model $f^*(x) = \sin(2\pi\frac{9}{70}x)$.
- (vi) *rbf*: Single RBF function $f^*(x) = \rho_1(x; 3, -10, 1)$.

The noise variance is set to $\sigma^2 = 0.01$, and assumed to be known. We set the number of RBF units to $H = 2$ for the singular model, and $H = 6$ for the regular model, so that both models have the same degrees of freedom, $D_{\text{sin-RBF}} = D_{\text{reg-RBF}} = 6$.

In Figure 13.6, we can observe the following: the singular RBF model can flexibly fit functions in different shapes (a) through (d), unless the function has too many peaks (e); the regular RBF model is not as flexible as the singular RBF model (a) through (d), unless the peaks and valleys match the predefined means of the RBF units (e). Actually, the frequency of *sin-alg* is aligned so that the peaks and the valleys match Eq. (13.112). These observations are quantitatively supported by the generalization error and the training error shown in Figure 13.7, leaving us an impression that the singular RBF model

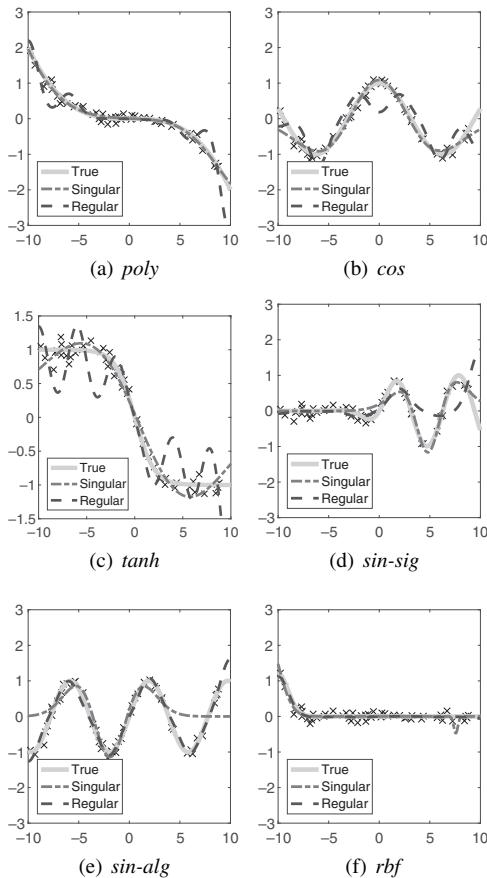


Figure 13.6 Trained regression functions by the singular RBF model (13.108) with $H = 2$ RBF units, and the regular RBF model (13.110) with $H = 6$ RBF units.

with two modifiable basis functions is more flexible than the regular RBF model with six prefixed basis functions.

However, *flexibility* is granted at the risk of overfitting to noise, which can be observed in Figure 13.6(f). We can see that the true RBF function at $x = -10$ is captured by both models. However, the singular RBF model shows a small valley around $x = 8$, which is a consequence of overfitting to sample noise. Figure 13.7 also shows that, in the *rbf* case, the singular RBF model gives lower training error and higher generalization error than the regular RBF model—typical behavior when overfitting occurs. This overfitting tendency is reflected to the generalization and the training coefficients.

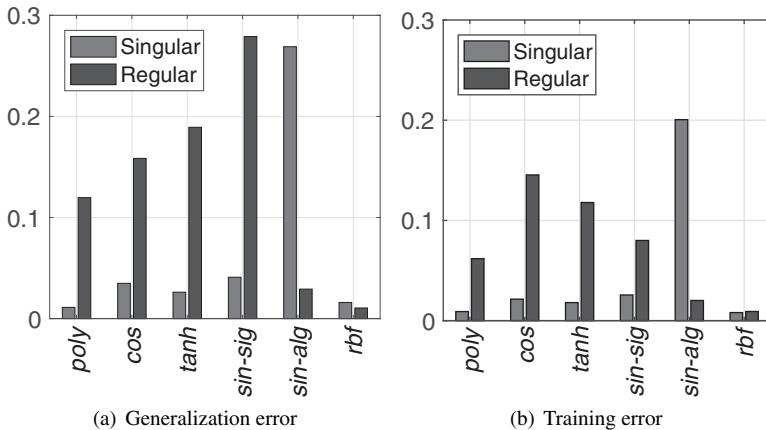


Figure 13.7 The generalization error and the training error by the singular RBF model and the regular RBF model.

Apparently, if the true function is realizable, i.e., $\exists \mathbf{w}^* \text{s.t. } f^*(x) = f(x; \mathbf{w}^*)$, the true distribution (13.114) is realizable by the RBF regression model (Eq. (13.108) or (13.110)). In the examples (a) through (e) in Figure 13.7, the true function is not realizable by the RBF regression model. In such cases, the generalization error and the training error do not converge to zero, and it holds that $\overline{\text{GE}}(N) = \Theta(1)$ and $\overline{\text{TE}}(N) = \Theta(1)$ for the best learning algorithm. On the other hand, the true function (f) consists of a single RBF unit, and furthermore its mean and variance match those of the first unit of the regular RBF model (see Eqs. (13.112) and (13.113)). Accordingly, the true function (f) and therefore the true distribution (13.114) in the example (f) are realizable, i.e., $\exists \mathbf{w}^*, \text{s.t. } q(y|x) = p(y|x, \mathbf{w}^*)$, by both of the singular RBF model (13.108) and the regular RBF model (13.110).

When the true parameter \mathbf{w}^* exists, the generalization error converges to zero, and, for any reasonable learning algorithm, the average generalization error and the average training error can be asymptotically expanded as Eqs. (13.15) and (13.16):

$$\begin{aligned}\overline{\text{GE}}(N) &= \lambda N^{-1} + o(N^{-1}), \\ \overline{\text{TE}}(N) &= \nu N^{-1} + o(N^{-1}).\end{aligned}$$

Since the regular RBF model (13.110) is regular, its generalization coefficient and the training coefficient are given by

$$2\lambda_{\text{reg-RBF}} = -2\nu_{\text{reg-RBF}} = D = H_{\text{reg-RBF}} \quad (13.115)$$

for ML learning, MAP learning, and Bayesian learning (under the additional regularity conditions on the prior). On the other hand, the generalization coefficient and the training coefficient for the singular RBF model (13.108) are unknown and can be significantly different from the regular models. As will be introduced in Section 13.5.3, the generalization coefficients of ML learning and MAP learning for various singular models have been clarified, and all results that have been found so far satisfy

$$2\lambda_{\text{Singular}}^{\text{ML}} \geq D, \quad 2\lambda_{\text{Singular}}^{\text{MAP}} \geq D, \quad (13.116)$$

where D is the parameter dimensionality. By comparing Eq. (13.116) with Eq. (13.115) (or Eqs. (13.67) and (13.81)), we find that the ML and the MAP generalization coefficients per single model parameter in singular models are larger than those in the regular models, which implies that singular models tend to overfit more than the regular models.

We can explain this phenomenon as an effect of the neighborhood structure around singularities. Recall the example (f), where the singular RBF model and the regular RBF model learn the true distribution $f^*(x) = \rho_1(x; 3, -10, 1)$. For the singular RBF model, $\mathbf{w}_{\text{sin-RBF}}^* = (a_1, a_2, b_1, b_2, c_1^2, c_2^2) = (3, 0, -10, *, 1, *)$, where $*$ allows any value in the domain, are possible true parameters, while, for the regular RBF model, $\mathbf{w}_{\text{sin-RBF}}^* = (a_1, a_2, \dots, a_6) = (3, 0, \dots, 0)$ is the unique true parameter. Figure 13.8(a) shows the space of the three parameters (a_2, b_2, c_2^2) of the second RBF unit of the singular RBF model, in which the true parameter is *on the singularities*. Since the true parameter extends over the two-dimensional half-space $\{(b_2, c_2^2); b_2 \in \mathbb{R}, c_2^2 \in \mathbb{R}_{++}\}$, the neighborhood (shown by small arrows) contains any RBF with adjustable mean and variance. Although the estimated parameter converges to the singularities in the asymptotic limit, ML learning on finite training samples tries to fit the noise, which contaminates the training samples, by *selecting the optimal basis function*, where the optimality is in terms of the training error. On the other hand, Figure 13.8(b) shows the parameter space $(a_h, \check{b}_h, \check{c}_h^2)$ for $h = 2, \dots, 4$. For each h , the true distribution corresponds to a single point, indicated by a shadowed circle, and its neighborhood extends only in one direction, i.e., $a_h = 0 \pm \varepsilon$ with a *prefixed* RBF basis specified by the constants $(\check{b}_h, \check{c}_h^2)$. Consequently, with the same number of redundant parameters as the singular RBF model, ML learning tries to fit the training noise only with those three basis functions.

Although the probability that the three prefixed basis functions can fit the training noise better than a single flexible basis function is not zero, we would expect that the singular RBF model would likely capture the noise more flexibly than the regular RBF model. This intuition is supported by previous theoretical work that showed Eq. (13.116) in many singular models,

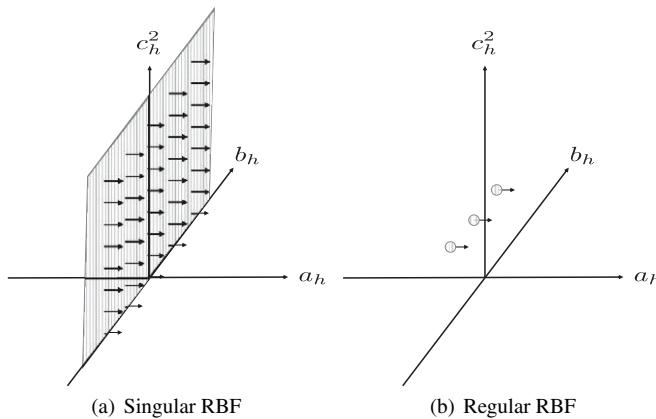


Figure 13.8 Neighborhood of the true distribution in the *rbf* example. (a) The parameter space of the second ($h = 2$) RBF unit of the singular RBF model. The true parameter is on the singularities, of which the neighborhood contains any RBF with adjustable mean and variance. (b) The parameter space of the second to the fourth ($h = 2, \dots, 4$) RBF units of the regular RBF model. With the same degrees of freedom as a single singular RBF unit, the neighborhood of the true parameter contains only three different RBF bases with *prefixed* means and variances.

as well as the numerical example in Figure 13.6. We call this phenomenon, i.e., singular models tending to overfit more than regular models, the *basis selection effect*. Although Eq. (13.116) was shown for ML learning and MAP learning, the basis selection effect should occur for any reasonable learning algorithms, including Bayesian learning. However, in Bayesian learning, this effect is canceled by the other effect of singularities, which is explained in the following subsection.

Integration Effect

Assume that, in Bayesian learning with a singular model, we adopt a prior distribution $p(\mathbf{w})$ bounded as $0 < p(\mathbf{w}) < \infty$ at any $\mathbf{w} \in \mathcal{W}$. This assumption is the same as one of the regularity conditions in Section 13.4.1. However, this assumption excludes the use of the Jeffreys prior (see Appendix B) and positive mass is distributed over singularities. As discussed in detail in Chapter 7, this prior choice leads to *nonuniformity of the volume element* and favors models with smaller degrees of freedom, if a learning algorithm involving integral computations in the parameter space is adopted. As a result, singularities induce MIR in Bayesian learning and its approximation methods, e.g., VB learning. Importantly, the integration effect does not occur in point estimation methods, including ML learning and MAP learning, since the nonuniformity of

the volume element affects the estimator only through *integral computations*. We call this phenomenon the *integration effect* of singularities.

The basis selection effect and the integration effect influence the learning behavior in the opposite way: the former intensifies overfitting, while the latter suppresses it. A question is which is stronger in Bayesian learning. Singular learning theory, which will be introduced in Section 13.5.4, has already answered this question. The following has been shown for *any* singular models:

$$2\lambda_{\text{Singular}}^{\text{Bayes}} \leq D. \quad (13.117)$$

Comparing Eq. (13.117) with Eq. (13.115) (or Eq. (13.94)), we find that the Bayes generalization coefficient per single model parameter in the singular models is smaller than that in the regular models. Note that the conclusion is opposite to ML learning and MAP learning—singular models overfit training noise more than the regular models in ML learning and MAP learning, while they less overfit in Bayesian learning. Since the basis selection effect should occur in any reasonable learning algorithm, we can interpret Eq. (13.117) as evidence that the integration effect is stronger than the basis selection effect in Bayesian learning.

One might wonder why Bayesian learning is not analyzed with the Jeffreys prior—the parameterization invariant noninformative prior. Actually, the Jeffreys prior, or other prior distribution with zero mass at the singularities, is rarely used in singular models because of the computational reasons: when the computational tractability relies on the (conditional) conjugacy, the Jeffreys prior is out of choice in singular models; when some sampling method is used for approximating the Bayes posterior, the diverging outskirts of the Jeffreys prior prevents the sampling sequence to converge. Note that this excludes the empirical Bayesian procedure, where the prior can be collapsed after training. Little is known about the learning behavior of empirical Bayesian learning in singular models, and the asymptotic learning theory part (Part IV) of this book also excludes this case.

In the following subsections, we give a brief summary of theoretical results that revealed learning properties of singular models.

13.5.2 Conditions Assumed in Asymptotic Theory for Singular Models

Singular models were analyzed under the following conditions on the true distribution and the prior distribution:

- (i) The true distribution is *realizable* by the statistical model, i.e.,
 $\exists \mathbf{w}^* \text{ s.t. } q(\mathbf{x}) = p(\mathbf{x}|\mathbf{w}^*)$.
- (ii) The prior $p(\mathbf{w})$ is twice differentiable and bounded as $0 < p(\mathbf{w}) < \infty$ at any $\mathbf{w} \in \mathcal{W}$.

Under the second condition, the prior choice does not affect the generalization coefficient. Accordingly, the results, introduced in the following subsection, for ML learning can be directly applied to MAP learning.

13.5.3 ML Learning and MAP Learning

Fukumizu (1999) analyzed the asymptotic behavior of the generalization error of ML learning for the reduced rank regression (RRR) model (3.36), by applying the *random matrix theory* to evaluate the singular value distribution of the ML estimator. Specifically, the large-scale limit where the dimensions of the input and the output are infinitely large was considered, and the exact generalization coefficient was derived. The training coefficient can be obtained in the same way (Nakajima and Watanabe, 2007).

The Gaussian mixture model (GMM) (4.6) has been studied as a prototype of singular models in the case of ML learning. Akaho and Kappen (2000) showed that the generalization error and the training error behave quite differently from regular models. As defined in Eq. (13.12), $-N \cdot \text{TE}^{\text{ML}}(X)$ is the log-likelihood ratio, which asymptotically follows the chi-squared distribution for regular models, while little is known about its behavior for singular models. In fact, it is conjectured for the spherical GMM (4.6) that the log-likelihood ratio diverges to infinity in the order of $\log \log N$ (Hartigan, 1985). For mixture models with discrete components such as binomial mixture models, the asymptotic distribution of the log-likelihood ratio was studied through the distribution of the maximum of the Gaussian random field (Bickel and Chernoff, 1993; Takemura and Kuriki, 1997; Kuriki and Takemura, 2001).

Based on the idea of locally conic parameterization (Dacunha-Castelle and Gassiat, 1997), the asymptotic behaviors of the log-likelihood ratio in some singular models were analyzed. For some mixture models with continuous components, including GMMs, it can be proved that the log-likelihood ratio diverges to infinity as $N \rightarrow \infty$. In neural networks, it is known that the log-likelihood ratio diverges in the order of $\log N$ when there are at least two redundant hidden units (Fukumizu, 2003; Hagiwara and Fukumizu, 2008).

In all previous works, the obtained generalization coefficient or its equivalent satisfies Eq. (13.116).

13.5.4 Singular Learning Theory for Bayesian Learning

For analyzing the generalization performance of Bayesian learning, a general approach, called the *singular learning theory (SLT)*, was established, based on the mathematical techniques in algebraic geometry (Watanabe, 2001a, 2009).

The average relative Bayes free energy (13.23),

$$\begin{aligned}\bar{F}^{\text{Bayes}}(N) &= \left\langle \log \frac{\prod_{n=1}^N q(\mathbf{x}^{(n)})}{\int p(\mathbf{w}) \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}) d\mathbf{w}} \right\rangle_{q(X)} \\ &= - \left\langle \log \int \exp \left(-N \log \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{w})} \right) \cdot p(\mathbf{w}) d\mathbf{w} \right\rangle_{q(\mathbf{x})},\end{aligned}$$

can be approximated as

$$\bar{F}^{\text{Bayes}}(N) \approx -\log \int \exp(-NE(\mathbf{w})) \cdot p(\mathbf{w}) d\mathbf{w}, \quad (13.118)$$

where

$$E(\mathbf{w}) = \left\langle \log \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{w})} \right\rangle_{q(\mathbf{x})} \quad (13.119)$$

is the KL divergence between the true distribution $q(\mathbf{x})$ and the model distribution $p(\mathbf{x}|\mathbf{w})$.⁸

Let us see the KL divergence (13.119) as the energy in physics, and define the *state density* function for the energy value $s > 0$:

$$v(s) = \int \delta(s - E(\mathbf{w})) \cdot p(\mathbf{w}) d\mathbf{w}, \quad (13.120)$$

where $\delta(\cdot)$ is the Dirac delta function (located at the origin). Note that the state density (13.120) and the (approximation to the relative) Bayes free energy (13.118) are connected by the Laplace transform:

$$\bar{F}^{\text{Bayes}}(N) = -\log \int \exp(-s)v\left(\frac{s}{N}\right) \frac{ds}{N}. \quad (13.121)$$

Define furthermore the *zeta function* as the Mellin transform, an extension of the Laplace transform, of the state density (13.120):

$$\zeta(z) = \int s^z v(s) ds = \int E(\mathbf{w})^z p(\mathbf{w}) d\mathbf{w}. \quad (13.122)$$

The zeta function (13.122) is a function of a complex number $z \in \mathbb{C}$, and it was proved that all the poles of $\zeta(z)$ are real, negative, and rational numbers.

⁸ It holds that $\bar{F}^{\text{Bayes}}(N) = -\log \int \exp(-NE(\mathbf{w})) \cdot p(\mathbf{w}) d\mathbf{w} + O(1)$ if the support of the prior is compact (Watanabe, 2001a, 2009).

By using the relations through Laplace/Mellin transform among the free energy (13.118), the state density (13.120), and the zeta function (13.122), Watanabe (2001a) proved the following theorem:

Theorem 13.13 (Watanabe, 2001a, 2009) *Let $0 > -\lambda_1 > -\lambda_2 > \dots$ be the sequence of the poles of the zeta function (13.122) in the decreasing order, and m_1, m_2, \dots be the corresponding orders of the poles. Then the average relative Bayes free energy (13.119) can be asymptotically expanded as*

$$\bar{F}^{\text{Bayes}}(N) = \lambda_1 \log N - (m_1 - 1) \log \log N + O(1). \quad (13.123)$$

Let $c(N) = \bar{F}^{\text{Bayes}}(N) - \lambda_1 \log N + (m_1 - 1) \log \log N$ be the $O(1)$ term in Eq. (13.123). The relation (13.24) between the generalization error and the free energy leads to the following corollary:

Corollary 13.14 (Watanabe, 2001a, 2009) *If $c(N+1) - c(N) = o\left(\frac{1}{N \log N}\right)$, the average generalization error (13.13) can be asymptotically expanded as*

$$\overline{\text{GE}}^{\text{Bayes}}(N) = \frac{\lambda_1}{N} - \frac{m_1 - 1}{N \log N} + o\left(\frac{1}{N \log N}\right). \quad (13.124)$$

To sum up, finding the maximum pole λ_1 of the zeta function $\zeta(z)$ gives the Bayes free energy coefficient

$$\lambda'^{\text{Bayes}} = \lambda_1,$$

which is equal to the Bayes generalization coefficient

$$\lambda^{\text{Bayes}} = \lambda_1.$$

Note that Theorem 13.13 and Corollary 13.14 hold both for regular and singular models. As discussed in Section 7.3.2, MIR (or the integration effect of singularities) is caused by strong nonuniformity of the density of the volume element. Since the state density (13.120) reflects the strength of the nonuniformity, one can see that finding the maximum pole of $\zeta(z)$ amounts to finding the strength of the nonuniformity at the most concentrated point.

Some general inequalities were proven (Watanabe, 2001b, 2009):

- If the prior is positive at any singular point, i.e., $p(\mathbf{w}) > 0$, $\forall \mathbf{w} \in \{\mathbf{w}; \det(\mathbf{F}(\mathbf{w})) = 0\}$, then

$$2\lambda'^{\text{Bayes}} = 2\lambda^{\text{Bayes}} \leq D. \quad (13.125)$$

- If the Jeffreys prior (see Appendix B.4) is adopted, for which $p(\mathbf{w}) = 0$ holds at any singular point, then

$$2\lambda'^{\text{Bayes}} = 2\lambda^{\text{Bayes}} \geq D. \quad (13.126)$$

Some cases have been found where $2\lambda'^{\text{Bayes}} = 2\lambda^{\text{Bayes}}$ are strictly larger than D .

These results support the discussion in Section 13.5.1 on the two effects of singularities: Eq. (13.125) implies that the integration effect dominates the basis selection effect in Bayesian learning, and Eq. (13.126) implies that the basis selection effect appears also in Bayesian learning if the integration effect is suppressed by using the Jeffreys prior.

Theorem 13.13 and Corollary 13.14 hold for general statistical models, while they do not immediately tell us learning properties of singular models. This is because finding the maximum pole of the zeta function $\zeta(z)$ is not an easy task, and requires a specific technique in algebraic geometry called the *resolution of singularities*. Good news is that, when any pole larger than $-D/2$ is found, it provides an upper bound of the generalization coefficient and thus guarantees the performance with a tighter bound (Theorem 13.13 implies that the larger the found pole is, the tighter the provided bound is).

For the RRR model (Aoyagi and Watanabe, 2005) and for the GMM (Aoyagi and Nagata, 2012), the maximum pole was found for general cases, and therefore the exact value of the free energy coefficient, as well as the generalization coefficient, was obtained. In other singular models, including neural networks (Watanabe, 2001a), mixture models (Yamazaki and Watanabe, 2003a), hidden Markov models (Yamazaki and Watanabe, 2005), and Bayesian networks (Yamazaki and Watanabe, 2003b; Rusakov and Geiger, 2005), upper-bounds of the free energy coefficient were obtained by finding some poles of the zeta function. An effort has been made to perform the resolution of singularities systematically by using the newton diagram (Yamazaki and Watanabe, 2004).

13.5.5 Information Criteria for Singular Models

The information criteria introduced in Section 13.4.8 rely on the learning theory under the regularity conditions. Therefore, although they were sometimes applied for model selection in singular models, their relations to generalization properties, e.g., AIC to the ML generalization error, and BIC to the Bayes free energy, do not generally hold. In the following, we introduce information criteria applicable for general statistical models including the regular and the singular models (Watanabe, 2009, 2010, 2013). They also cover a generalization of Bayesian learning.

Consider a learning method, called *generalized Bayesian learning*, based on the *generalized posterior distribution*,

$$p^{(\beta)}(\mathbf{w}|\mathbf{X}) = \frac{p(\mathbf{w}) \prod_{n=1}^N \{p(\mathbf{x}^{(n)}|\mathbf{w})\}^\beta}{\int p(\mathbf{w}) \prod_{n=1}^N \{p(\mathbf{x}^{(n)}|\mathbf{w})\}^\beta d\mathbf{w}}, \quad (13.127)$$

where β , called the *inverse temperature parameter*, modifies the importance of the likelihood per training sample. The prediction is made by the *generalized predictive distribution*,

$$p^{(\beta)}(\mathbf{x}|\mathbf{X}) = \langle p(\mathbf{x}|\mathbf{w}) \rangle_{p^{(\beta)}(\mathbf{w}|\mathbf{X})}. \quad (13.128)$$

Generalized Bayesian learning covers both Bayesian learning and ML learning as special cases: when $\beta = 1$, the generalized posterior distribution (13.127) is reduced to the Bayes posterior distribution (13.9), with which the generalized predictive distribution (13.128) gives the Bayes predictive distribution (13.8); As β increases, the probability mass of the generalized posterior distribution concentrates around the ML estimator, and, in the limit when $\beta \rightarrow \infty$, the generalized predictive distribution converges to the ML predictive distribution (13.6).

Define the following quantities:

$$\text{GL}(\mathbf{X}) = - \left\langle \log \int p(\mathbf{x}|\mathbf{w}) p^{(\beta)}(\mathbf{w}|\mathbf{X}) d\mathbf{w} \right\rangle_{q(\mathbf{x})}, \quad (13.129)$$

$$\text{TL}(\mathbf{X}) = - \frac{1}{N} \sum_{n=1}^N \log \int p(\mathbf{x}^{(n)}|\mathbf{w}) p^{(\beta)}(\mathbf{w}|\mathbf{X}) d\mathbf{w}, \quad (13.130)$$

$$\text{GGL}(\mathbf{X}) = - \left\langle \int (\log p(\mathbf{x}|\mathbf{w})) p^{(\beta)}(\mathbf{w}|\mathbf{X}) d\mathbf{w} \right\rangle_{q(\mathbf{x})}, \quad (13.131)$$

$$\text{GTL}(\mathbf{X}) = - \frac{1}{N} \sum_{n=1}^N \int (\log p(\mathbf{x}^{(n)}|\mathbf{w})) p^{(\beta)}(\mathbf{w}|\mathbf{X}) d\mathbf{w}, \quad (13.132)$$

which are called the *Bayes generalization loss*, the *Bayes training loss*, the *Gibbs generalization loss*, and the *Gibbs training loss*, respectively. The generalization error and the training error of generalized Bayesian learning are, respectively, related to the Bayes generalization loss and the Bayes training loss as follows (Watanabe, 2009):

$$\begin{aligned} \text{GE}^{(\beta)}(\mathbf{X}) &= \left\langle \log \frac{q(\mathbf{x})}{\int p(\mathbf{x}|\mathbf{w}) p^{(\beta)}(\mathbf{w}|\mathbf{X}) d\mathbf{w}} \right\rangle_{q(\mathbf{x})} \\ &= \text{GL}(\mathbf{X}) - S, \end{aligned} \quad (13.133)$$

$$\begin{aligned} \text{TE}^{(\beta)}(\mathbf{X}) &= \frac{1}{N} \sum_{n=1}^N \log \frac{q(\mathbf{x}^{(n)})}{\int p(\mathbf{x}^{(n)}|\mathbf{w}) p^{(\beta)}(\mathbf{w}|\mathbf{X}) d\mathbf{w}} \\ &= \text{TL}(\mathbf{X}) - S_N(\mathbf{X}), \end{aligned} \quad (13.134)$$

where

$$S = -\langle \log q(\mathbf{x}) \rangle_{q(\mathbf{x})} \quad \text{and} \quad S_N(\mathbf{X}) = -\frac{1}{N} \sum_{n=1}^N \log q(\mathbf{x}^{(n)}) \quad (13.135)$$

are the *entropy* of the true distribution and its empirical version, respectively.⁹ Also, Gibbs counterparts have the following relations:

$$\begin{aligned} \text{GGE}^{(\beta)}(\mathbf{X}) &= \left\langle \int \left(\log \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{w})} \right) p^{(\beta)}(\mathbf{w}|\mathbf{X}) d\mathbf{w} \right\rangle_{q(\mathbf{x})} \\ &= \text{GGL}(\mathbf{X}) - S, \end{aligned} \quad (13.136)$$

$$\begin{aligned} \text{GTE}^{(\beta)}(\mathbf{X}) &= \frac{1}{N} \sum_{n=1}^N \int \left(\log \frac{q(\mathbf{x}^{(n)})}{p(\mathbf{x}^{(n)}|\mathbf{w})} \right) p^{(\beta)}(\mathbf{w}|\mathbf{X}) d\mathbf{w} \\ &= \text{GTL}(\mathbf{X}) - S_N(\mathbf{X}). \end{aligned} \quad (13.137)$$

Here $\text{GGE}^{(\beta)}(\mathbf{X})$ and $\text{GTE}^{(\beta)}(\mathbf{X})$ are the generalization error and the training error, respectively, of *Gibbs learning*, where prediction is made by $p(\mathbf{x}|\mathbf{w})$ with its parameter \mathbf{w} sampled from the generalized posterior distribution (13.127).

The following relations were proven (Watanabe, 2009):

$$\langle \text{GL}(\mathbf{X}) \rangle_{q(\mathbf{X})} = \langle \text{TL}(\mathbf{X}) \rangle_{q(\mathbf{X})} + 2\beta \left(\langle \text{GTL}(\mathbf{X}) \rangle_{q(\mathbf{X})} - \langle \text{TL}(\mathbf{X}) \rangle_{q(\mathbf{X})} \right) + o(N^{-1}), \quad (13.138)$$

$$\langle \text{GGL}(\mathbf{X}) \rangle_{q(\mathbf{X})} = \langle \text{GTL}(\mathbf{X}) \rangle_{q(\mathbf{X})} + 2\beta \left(\langle \text{GTL}(\mathbf{X}) \rangle_{q(\mathbf{X})} - \langle \text{TL}(\mathbf{X}) \rangle_{q(\mathbf{X})} \right) + o(N^{-1}), \quad (13.139)$$

which imply that asymptotically unbiased estimators for generalization losses (the left-hand sides of Eqs. (13.138) and (13.139)) can be constructed from training losses (the right-hand sides). The aforementioned equations lead to *widely applicable information criteria (WAIC)* (Watanabe, 2009), defined as

$$\text{WAIC}_1 = \text{TL}(\mathbf{X}) + 2\beta (\text{GTL}(\mathbf{X}) - \text{TL}(\mathbf{X})), \quad (13.140)$$

$$\text{WAIC}_2 = \text{GTL}(\mathbf{X}) + 2\beta (\text{GTL}(\mathbf{X}) - \text{TL}(\mathbf{X})). \quad (13.141)$$

⁹ $S_N(\mathbf{X})$ was defined in Eq. (13.20), and it holds that $S = \langle S_N(\mathbf{X}) \rangle_{q(\mathbf{X})}$.

Clearly, WAIC_1 and WAIC_2 are asymptotically unbiased estimators for the Bayes generalization loss $\text{GL}(\mathbf{X})$ and the Gibbs generalization loss $\text{GGL}(\mathbf{X})$, respectively, and therefore minimizing them amounts to minimizing the Bayes generalization error (13.133) and the Gibbs generalization error (13.136), respectively.

The training losses, $\text{TL}(\mathbf{X})$ and $\text{GTL}(\mathbf{X})$, can be computed by, e.g., MCMC sampling (see Sections 2.2.4 and 2.2.5). Let $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(T)}$ be samples drawn from the generalized posterior distribution (13.127). Then we can estimate the training losses by

$$\text{TL}(\mathbf{X}) \approx -\frac{1}{N} \sum_{n=1}^N \log \left(\frac{1}{T} \sum_{t=1}^T p(\mathbf{x}^{(n)} | \mathbf{w}^{(t)}) \right), \quad (13.142)$$

$$\text{GTL}(\mathbf{X}) \approx -\frac{1}{N} \sum_{n=1}^N \frac{1}{T} \sum_{t=1}^T \log p(\mathbf{x}^{(n)} | \mathbf{w}^{(t)}). \quad (13.143)$$

WAIC can be seen as an extension of AIC, since minimizing it amounts to minimizing an asymptotically unbiased estimator for the generalization error. Indeed, under the regularity conditions, it holds that

$$\lim_{\beta \rightarrow \infty} 2\beta (\text{GTL}(\mathbf{X}) - \text{TL}(\mathbf{X})) = \frac{D}{N},$$

and therefore

$$\text{WAIC}_1, \text{WAIC}_2 \rightarrow \frac{\text{AIC}}{2N} \quad \text{as} \quad \beta \rightarrow \infty.$$

An extension of BIC was also proposed. The widely applicable Bayesian information criterion (WBIC) (Watanabe, 2013) is defined as

$$\text{WBIC} = - \sum_{n=1}^N \int \log p(\mathbf{x}^{(n)} | \mathbf{w}) p^{(\beta=1/\log N)}(\mathbf{w} | \mathbf{X}) d\mathbf{w}, \quad (13.144)$$

where $p^{(\beta=1/\log N)}(\mathbf{w} | \mathbf{X})$ is the generalized posterior distribution (13.127) with the inverse temperature parameter set to $\beta = 1/\log N$. It was shown that

$$F^{\text{Bayes}}(\mathbf{X}) \left(\equiv -\log \int p(\mathbf{w}) \prod_{n=1}^N p(\mathbf{x}^{(n)} | \mathbf{w}) d\mathbf{w} \right) = \text{WBIC} + O_p(\sqrt{\log N}), \quad (13.145)$$

and therefore WBIC can be used as an estimator or approximation for the Bayes free energy (13.18) when N is large. It was also shown that, under the regularity conditions, it holds that

$$\text{WBIC} = \frac{\text{BIC}}{2} + O_p(1).$$

WBIC (13.144) can be estimated, similarly to WAIC, from samples $\mathbf{w}_{\beta=1/\log N}^{(1)}, \dots, \mathbf{w}_{\beta=1/\log N}^{(T)}$ drawn from $p^{(\beta=1/\log N)}(\mathbf{w}|\mathbf{X})$ as

$$\text{WBIC} \approx - \sum_{n=1}^N \frac{1}{T} \sum_{t=1}^T \log p(\mathbf{x}^{(n)} | \mathbf{w}_{\beta=1/\log N}^{(t)}). \quad (13.146)$$

Note that evaluating the Bayes free energy (13.18) is much more computationally demanding in general. For example, the *all temperatures method* (Watanabe, 2013) requires posterior samples $\{\mathbf{w}_{\beta_j}^{(t)}\}$ for many $0 = \beta_1 < \beta_2 < \dots < \beta_J = 1$, and estimates the Bayes free energy as

$$F^{\text{Bayes}}(\mathbf{X}) \approx - \sum_{j=1}^{J-1} \log \frac{1}{T_j} \sum_{t=1}^{T_j} \exp \left((\beta_{j+1} - \beta_j) \sum_{n=1}^N \log p(\mathbf{x}^{(n)} | \mathbf{w}_{\beta_j}^{(t)}) \right).$$

13.6 Asymptotic Learning Theory for VB Learning

In the rest of Part IV, we describe asymptotic learning theory for VB learning in detail. Here we give an overview of the subsequent chapters.

VB learning is rarely applied to regular models.¹⁰ Actually, if the model (and the prior) satisfies the regularity conditions, Laplace approximation (2.2.1) can give a good approximation to the posterior, because of the asymptotic normality (Theorem 13.3). Accordingly, we focus on singular models when analyzing VB learning.

We are interested in generalization properties of the VB posterior, which is defined as

$$\widehat{\mathbf{r}} \equiv \underset{\mathbf{r}}{\operatorname{argmin}} F(\mathbf{r}), \quad \text{s.t.} \quad \mathbf{r} \in \mathcal{G}, \quad (13.147)$$

where

$$F(\mathbf{r}) = \left\langle \log \frac{r(\mathbf{w})}{p(\mathbf{w}) \prod_{n=1}^N p(\mathbf{x}^{(n)} | \mathbf{w})} \right\rangle_{r(\mathbf{w})} = \text{KL}(r(\mathbf{w}) \| p(\mathbf{w} | \mathbf{X})) + F^{\text{Bayes}}(\mathbf{X}) \quad (13.148)$$

is the free energy and \mathcal{G} is the model-specific constraint, imposed for computational tractability, on the approximate posterior.

¹⁰ VB learning is often applied to a linear model with an ARD prior. In such a model, the model likelihood satisfies the regularity conditions, while the prior does not. Actually, the model exhibits characteristics of singular models, since it can be translated to a singular model with a constant prior (see Section 7.5). In Part IV, we only consider the case where the prior is fixed, without any hyperparameter optimized.

With the VB predictive distribution

$$p^{\text{VB}}(\mathbf{x}|\mathbf{X}) = \langle p(\mathbf{x}|\mathbf{w}) \rangle_{\widehat{r}(\mathbf{w})},$$

the generalization error (13.10) and the training error (13.11) are defined and analyzed.

We also analyze the *VB free energy*,

$$F^{\text{VB}}(\mathbf{X}) = F(\widehat{r}) = \min_r F(r). \quad (13.149)$$

Since Eq. (13.148) implies that

$$F^{\text{VB}}(\mathbf{X}) - F^{\text{Bayes}}(\mathbf{X}) = \text{KL}(\widehat{r}(\mathbf{w})||p(\mathbf{w}|\mathbf{X})),$$

comparing the VB free energy and the Bayes free energy reveals how accurately VB learning approximates Bayesian learning.

Similarly to the analysis of Bayesian learning, we investigate the asymptotic behavior of the *relative VB free energy*,

$$\widetilde{F}^{\text{VB}}(\mathbf{X}) = F^{\text{VB}}(\mathbf{X}) - NS_N(\mathbf{X}) = \lambda'^{\text{VB}} \log N + o_p(\log N), \quad (13.150)$$

where $S_N(\mathbf{X})$ is the empirical entropy defined in Eq. (13.20), and λ'^{VB} is called the VB free energy coefficient.

Chapter 14 introduces asymptotic VB theory for the RRR model. This model was relatively easily analyzed by using the analytic-form solution for fully observed matrix factorization (Chapter 6), and the exact values of the VB generalization coefficient, the VB training coefficient, and the VB free energy coefficient were derived (Nakajima and Watanabe, 2007). Since generalization properties of ML learning and Bayesian learning have also been clarified (Fukumizu, 1999; Aoyagi and Watanabe, 2005), similarities and dissimilarities among ML (and MAP) learning, Bayesian learning, and VB learning will be discussed.

Chapters 15 through 17 are devoted to asymptotic VB theory for latent variable models. Chapter 15 analyzes the VB free energy of mixture models. The VB free energy coefficients and their dependencies on prior hyperparameters are revealed. Chapter 16 proceeds to such analyses of the VB free energy for other latent variable models, namely, Bayesian networks, hidden Markov models, probabilistic context free grammar, and latent Dirichlet allocation. Chapter 17 provides a formula for general latent variable models, which reduces the asymptotic analysis of the VB free energy to that of the Bayes free energy introduced in Section 13.5.4. Those results will clarify phase transition phenomena with respect to the hyperparameter setting—the shape of the posterior distribution in the asymptotic limit drastically changes when some

hyperparameter value exceeds a certain threshold. Such implication suggests to practitioners how to choose hyperparameters.

Note that the relation (13.25) does not necessarily hold for VB learning and other approximate Bayesian methods, since Eq. (13.24) only holds for the exact Bayes predictive distribution. Therefore, unlike Bayesian learning, the asymptotic behavior of the VB free energy does not necessarily inform us of the asymptotic behavior of the VB generalization error. An effort on relating the VB free energy and the VB generalization error is introduced in Chapter 17, although clarifying VB generalization error requires further effort and techniques.

14

Asymptotic VB Theory of Reduced Rank Regression

In this chapter, we introduce asymptotic theory of VB learning in the reduced rank regression (RRR) model (Nakajima and Watanabe, 2007). Among the singular models, the RRR model is one of the simplest, and many aspects of its learning behavior have been clarified. Accordingly, we can discuss similarities and dissimilarities of ML (and MAP) learning, Bayesian learning, and VB learning in terms of generalization error, training error, and free energy. After defining the problem setting, we show theoretical results and summarize insights into VB learning that the analysis on the RRR model provides.

14.1 Reduced Rank Regression

RRR (Baldi and Hornik, 1995; Reinsel and Velu, 1998), introduced in Section 3.1.2 as a special case of fully observed matrix factorization, is a regression model with a rank- H ($\leq \min(L, M)$) linear mapping between input $\mathbf{x} \in \mathbb{R}^M$ and output $\mathbf{y} \in \mathbb{R}^L$:

$$\mathbf{y} = \mathbf{B}\mathbf{A}^\top \mathbf{x} + \boldsymbol{\varepsilon}, \quad (14.1)$$

where $\mathbf{A} \in \mathbb{R}^{M \times H}$ and $\mathbf{B} \in \mathbb{R}^{L \times H}$ are parameters to be estimated, and $\boldsymbol{\varepsilon}$ is observation noise. Assuming Gaussian noise $\boldsymbol{\varepsilon} \sim \text{Gauss}_L(\mathbf{0}, \sigma'^2 \mathbf{I}_L)$, the model distribution is given as

$$p(\mathbf{y}|\mathbf{x}, \mathbf{A}, \mathbf{B}) = (2\pi\sigma'^2)^{-L/2} \exp\left(-\frac{1}{2\sigma'^2} \|\mathbf{y} - \mathbf{B}\mathbf{A}^\top \mathbf{x}\|^2\right). \quad (14.2)$$

RRR is also called a *linear neural network*, since the three-layer neural network (7.13) is reduced to RRR (14.1) if the activation function $\psi(\cdot)$ is linear

(see also Figure 3.1). We assume conditionally conjugate Gaussian priors for the parameters:

$$p(\mathbf{A}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{A}\mathbf{C}_A^{-1}\mathbf{A}^\top)\right), \quad p(\mathbf{B}) \propto \exp\left(-\frac{1}{2}\text{tr}(\mathbf{B}\mathbf{C}_B^{-1}\mathbf{B}^\top)\right), \quad (14.3)$$

with diagonal covariances \mathbf{C}_A and \mathbf{C}_B :

$$\mathbf{C}_A = \mathbf{Diag}(c_{a_1}^2, \dots, c_{a_H}^2), \quad \mathbf{C}_B = \mathbf{Diag}(c_{b_1}^2, \dots, c_{b_H}^2),$$

for $c_{a_h}, c_{b_h} > 0, h = 1, \dots, H$. In the asymptotic analysis, we assume that the hyperparameters $\{c_{a_h}^2, c_{b_h}^2\}, \sigma'^2$ are fixed constants of the order of 1, i.e., $\{c_{a_h}^2, c_{b_h}^2\}, \sigma'^2 \sim \Theta(1)$ when $N \rightarrow \infty$.

The degree of freedom of the RRR model is, in general, different from the apparent number, $(M + L)H$, of entries of the parameters \mathbf{A} and \mathbf{B} . This is because of the trivial redundancy in parameterization—the transformation $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{A}\mathbf{T}^\top, \mathbf{B}\mathbf{T}^{-1})$ does not change the linear mapping $\mathbf{B}\mathbf{A}^\top$ for any nonsingular matrix $\mathbf{T} \in \mathbb{R}^{H \times H}$. Accordingly, the *essential* parameter dimensionality is counted as

$$D = H(M + L) - H^2. \quad (14.4)$$

Suppose we are given N training samples:

$$\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}); \mathbf{x}^{(n)} \in \mathbb{R}^M, \mathbf{y}^{(n)} \in \mathbb{R}^L, n = 1, \dots, N\}, \quad (14.5)$$

which are independently drawn from the true distribution $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}|\mathbf{x})q(\mathbf{x})$. We also use the matrix forms that summarize the inputs and the outputs separately:

$$\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})^\top \in \mathbb{R}^{N \times M}, \quad \mathbf{Y} = (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)})^\top \in \mathbb{R}^{N \times L}.$$

We suppose that the data are preprocessed so that the input and the output are centered, i.e.,

$$\frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)} = \mathbf{0} \quad \text{and} \quad \frac{1}{N} \sum_{n=1}^N \mathbf{y}^{(n)} = \mathbf{0}, \quad (14.6)$$

and the input is *prewhitened* (Hyvärinen et al., 2001), i.e.,

$$\frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)} \mathbf{x}^{(n)\top} = \frac{1}{N} \mathbf{X}^\top \mathbf{X} = \mathbf{I}_M. \quad (14.7)$$

The likelihood of the RRR model (14.1) on the training samples $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ is expressed as

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{1}{2\sigma'^2} \sum_{n=1}^N \|\mathbf{y}^{(n)} - \mathbf{B}\mathbf{A}^\top \mathbf{x}^{(n)}\|^2\right). \quad (14.8)$$

As shown in Section 3.1.2, the logarithm of the likelihood (14.8) can be written, as a function of the parameters, as follows:

$$\log p(\mathbf{Y}|\mathbf{X}, \mathbf{A}, \mathbf{B}) = -\frac{N}{2\sigma'^2} \|\mathbf{V} - \mathbf{B}\mathbf{A}^\top\|_{\text{Fro}}^2 + \text{const.}, \quad (14.9)$$

where

$$\mathbf{V} = \frac{1}{N} \sum_{n=1}^N \mathbf{y}^{(n)} \mathbf{x}^{(n)\top} = \frac{1}{N} \mathbf{Y}^\top \mathbf{X}. \quad (14.10)$$

Note that, unlike in Section 3.1.2, we here do not use the rescaled noise variance $\sigma^2 = \sigma'^2/N$, in order to make the dependence on the number N of samples clear for asymptotic analysis. Because the log-likelihood (14.9) is in the same form as that of the fully observed matrix factorization (MF), we can use the global VB solution, derived in Chapter 6, of the MF model for analyzing VB learning in the RRR model.

14.1.1 VB Learning

VB learning solves the following problem:

$$\widehat{r} = \underset{r}{\operatorname{argmin}} F(r) \quad \text{s.t.} \quad r(\mathbf{A}, \mathbf{B}) = r_A(\mathbf{A})r_B(\mathbf{B}), \quad (14.11)$$

where

$$F = \left\langle \log \frac{r_A(\mathbf{A})r_B(\mathbf{B})}{p(\mathbf{Y}|\mathbf{X}, \mathbf{A}, \mathbf{B})p(\mathbf{A})p(\mathbf{B})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}$$

is the free energy. As derived in Section 3.1, the solution to the problem (14.11) is in the following forms:

$$r_A(\mathbf{A}) = \text{MGauss}_{M,H}(\mathbf{A}; \widehat{\mathbf{A}}, \mathbf{I}_M \otimes \widehat{\Sigma}_A) \propto \exp \left\{ -\frac{\text{tr}((\mathbf{A} - \widehat{\mathbf{A}})\widehat{\Sigma}_A^{-1}(\mathbf{A} - \widehat{\mathbf{A}})^\top)}{2} \right\}, \quad (14.12)$$

$$r_B(\mathbf{B}) = \text{MGauss}_{L,H}(\mathbf{B}; \widehat{\mathbf{B}}, \mathbf{I}_L \otimes \widehat{\Sigma}_B) \propto \exp \left\{ -\frac{\text{tr}((\mathbf{B} - \widehat{\mathbf{B}})\widehat{\Sigma}_B^{-1}(\mathbf{B} - \widehat{\mathbf{B}})^\top)}{2} \right\}. \quad (14.13)$$

With the variational parameters $(\widehat{\mathbf{A}}, \widehat{\Sigma}_A, \widehat{\mathbf{B}}, \widehat{\Sigma}_B)$, the free energy can be explicitly written as

$$\begin{aligned}
2F = & NL \log(2\pi\sigma'^2) + \frac{\sum_{n=1}^N \|\mathbf{y}^{(n)}\|^2 - N \|\mathbf{V}\|_{\text{Fro}}^2}{\sigma'^2} \\
& + \frac{N \|\mathbf{V} - \widehat{\mathbf{B}}\mathbf{A}^\top\|_{\text{Fro}}^2}{\sigma'^2} + M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + L \log \frac{\det(\mathbf{C}_B)}{\det(\widehat{\Sigma}_B)} \\
& - (L+M)H + \text{tr} \left\{ \mathbf{C}_A^{-1} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) + \mathbf{C}_B^{-1} \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B \right) \right. \\
& \left. + N \sigma'^{-2} \left(-\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B \right) \right) \right\}. \quad (14.14)
\end{aligned}$$

We can further apply Corollary 6.6, which states that the VB learning problem (14.11) is decomposable in the following way. Let

$$\mathbf{V} = \sum_{h=1}^L \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top \quad (14.15)$$

be the singular value decomposition (SVD) of \mathbf{V} (defined in Eq. (14.10)), where γ_h (≥ 0) is the h th largest singular value, and $\boldsymbol{\omega}_{a_h}$ and $\boldsymbol{\omega}_{b_h}$ are the associated right and left singular vectors. Then the solution (or its equivalent) of the variational parameters $\widehat{\mathbf{A}} = (\widehat{\mathbf{a}}_1, \dots, \widehat{\mathbf{a}}_H)$, $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \dots, \widehat{\mathbf{b}}_H)$, $\widehat{\Sigma}_A$, $\widehat{\Sigma}_B$, which minimizes the free energy (14.14), can be expressed as follows:

$$\begin{aligned}
\widehat{\mathbf{a}}_h &= \widehat{\mathbf{a}}_h \boldsymbol{\omega}_{a_h}, & \widehat{\mathbf{b}}_h &= \widehat{\mathbf{b}}_h \boldsymbol{\omega}_{b_h}, \\
\widehat{\Sigma}_A &= \mathbf{Diag}(\widehat{\sigma}_{a_1}^2, \dots, \widehat{\sigma}_{a_H}^2), & \widehat{\Sigma}_B &= \mathbf{Diag}(\widehat{\sigma}_{b_1}^2, \dots, \widehat{\sigma}_{b_H}^2),
\end{aligned}$$

where $\{\widehat{\mathbf{a}}_h, \widehat{\mathbf{b}}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H$ are a new set of variational parameters. Thus, the VB posteriors (14.12) and (14.13) can be written as

$$r_A(\mathbf{A}) = \prod_{h=1}^H \text{Gauss}_M(\mathbf{a}_h; \widehat{\mathbf{a}}_h \boldsymbol{\omega}_{a_h}, \widehat{\sigma}_{a_h}^2 \mathbf{I}_M), \quad (14.16)$$

$$r_B(\mathbf{B}) = \prod_{h=1}^H \text{Gauss}_L(\mathbf{b}_h; \widehat{\mathbf{b}}_h \boldsymbol{\omega}_{b_h}, \widehat{\sigma}_{b_h}^2 \mathbf{I}_L), \quad (14.17)$$

with $\{\widehat{\mathbf{a}}_h, \widehat{\mathbf{b}}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2\}_{h=1}^H$ that are the solution of the following minimization problem:

$$\begin{aligned}
\text{Given } & \sigma'^2 \in \mathbb{R}_{++}, \quad \{c_{a_h}^2, c_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H, \\
& \min_{\{\widehat{\mathbf{a}}_h, \widehat{\mathbf{b}}_h, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2\}_{h=1}^H} F \\
\text{s.t. } & \{\widehat{\mathbf{a}}_h, \widehat{\mathbf{b}}_h \in \mathbb{R}, \widehat{\sigma}_{a_h}^2, \widehat{\sigma}_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H. \quad (14.18)
\end{aligned}$$

Here F is the free energy (14.14), which can be decomposed as

$$2F = NL \log(2\pi\sigma'^2) + \frac{\sum_{n=1}^N \|y^{(n)}\|^2}{\sigma'^2} + \sum_{h=1}^H 2F_h, \quad (14.19)$$

$$\text{where } 2F_h = M \log \frac{c_{a_h}^2}{\widehat{\sigma}_{a_h}^2} + L \log \frac{c_{b_h}^2}{\widehat{\sigma}_{b_h}^2} + \frac{\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2}{c_{b_h}^2} - (L+M) + \frac{N}{\sigma'^2} \left(-2\widehat{a}_h\widehat{b}_h\gamma_h + (\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2)(\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2) \right). \quad (14.20)$$

14.1.2 VB Solution

Let us derive an asymptotic-form VB solution from the nonasymptotic global VB solution, derived in Section 6. Theorem 6.7 leads to the following theorem:

Theorem 14.1 *The VB estimator $\widehat{\mathbf{U}}^{\text{VB}} \equiv \langle \mathbf{B}\mathbf{A}^\top \rangle_{r_A(A)r_B(B)}$ for the linear mapping of the RRR model (14.2) and (14.3) can be written as*

$$\widehat{\mathbf{U}}^{\text{VB}} = \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top = \sum_{h=1}^H \widehat{\gamma}_h^{\text{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \quad \text{where } \widehat{\gamma}_h^{\text{VB}} = \max(0, \check{\gamma}_h^{\text{VB}}) \quad (14.21)$$

for

$$\check{\gamma}_h^{\text{VB}} = \gamma_h \left(1 - \frac{\max(L, M)\sigma'^2}{N\gamma_h^2} \right) + O_p(N^{-1}). \quad (14.22)$$

For each component h , $\widehat{\gamma}_h^{\text{VB}} > 0$ if and only if $\gamma_h > \underline{\gamma}_h^{\text{VB}}$ for

$$\underline{\gamma}_h^{\text{VB}} = \sigma' \sqrt{\frac{\max(L, M)}{N}} + O(N^{-1}). \quad (14.23)$$

Proof Noting that Theorem 6.7 gives the VB solution for either \mathbf{V} or $\mathbf{V}^\top \in \mathbb{R}^{L \times M}$ that satisfies $L \leq M$, that the shrinkage estimator $\check{\gamma}_h^{\text{VB}}$ (given by Eq. (6.50)) is an increasing function of γ_h , and that $\check{\gamma}_h^{\text{VB}} = 0$ when γ_h is equal to the threshold $\underline{\gamma}_h^{\text{VB}}$ (given by Eq. (6.49)), we have Eq. (14.21) with

$$\begin{aligned} \check{\gamma}_h^{\text{VB}} &= \gamma_h \left(1 - \frac{\sigma'^2}{2N\gamma_h^2} \left(L + M + \sqrt{(M-L)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}} \right) \right) \\ &= \gamma_h \left(1 - \frac{\sigma'^2}{2N\gamma_h^2} \left(L + M + \sqrt{(M-L)^2 + O(\gamma_h^2)} \right) \right) \end{aligned}$$

$$\begin{aligned}
&= \begin{cases} \gamma_h \left(1 - \frac{\sigma'^2}{2N\gamma_h^2} (L + M + \max(L, M) - \min(L, M) + O(\gamma_h^2)) \right) & (\text{if } L \neq M) \\ \gamma_h \left(1 - \frac{\sigma'^2}{2N\gamma_h^2} (L + M + O(\gamma_h)) \right) & (\text{if } L = M) \end{cases} \\
&= \gamma_h \left(1 - \frac{\max(L, M)\sigma'^2}{N\gamma_h^2} (1 + O_p(\gamma_h)) \right) \\
&= \gamma_h \left(1 - \frac{\max(L, M)\sigma'^2}{N\gamma_h^2} \right) + O_p(N^{-1}),
\end{aligned} \tag{14.24}$$

and

$$\begin{aligned}
\underline{\gamma}_h^{\text{VB}} &= \frac{\sigma'}{\sqrt{N}} \sqrt{\frac{(L+M)}{2} + \frac{\sigma'^2}{2Nc_{a_h}^2 c_{b_h}^2} + \sqrt{\left(\frac{(L+M)}{2} + \frac{\sigma'^2}{2Nc_{a_h}^2 c_{b_h}^2} \right)^2 - LM}} \\
&= \frac{\sigma'}{\sqrt{N}} \sqrt{\frac{(L+M)}{2} + \frac{\sigma'^2}{2Nc_{a_h}^2 c_{b_h}^2} + \sqrt{\left(\frac{\max(L, M) - \min(L, M)}{2} \right)^2 + O(N^{-1})}} \\
&= \begin{cases} \frac{\sigma'}{\sqrt{N}} \sqrt{\frac{(L+M)}{2} + \frac{\sigma'^2}{2Nc_{a_h}^2 c_{b_h}^2} + \frac{\max(L, M) - \min(L, M)}{2} + O(N^{-1})} & (\text{if } L \neq M) \\ \frac{\sigma'}{\sqrt{N}} \sqrt{\frac{(L+M)}{2} + \frac{\sigma'^2}{2Nc_{a_h}^2 c_{b_h}^2} + O(N^{-1/2})} & (\text{if } L = M) \end{cases} \\
&= \frac{\sigma'}{\sqrt{N}} \sqrt{\max(L, M)} + O(N^{-1}),
\end{aligned}$$

which completes the proof. Note that we used $\gamma_h = O_p(1)$ to get Eq. (14.24). \square

Theorem 14.1 states that the VB estimator converges to the positive-part James–Stein (PJS) estimator (see Appendix A)—the same solution (Corollary 7.1) as the nonasymptotic MF solution with the flat prior. This is natural because the influence from the constant prior disappears in the asymptotic limit, making MAP learning converge to ML learning.

Corollary 6.8 leads to the following corollary:

Corollary 14.2 *The VB posterior of the RRR model (14.2) and (14.3) is given by Eqs. (14.16) and (14.17) with the variational parameters given as follows: if $\gamma_h > \underline{\gamma}_h^{\text{VB}}$,*

$$\begin{aligned}
\widehat{a}_h &= \pm \sqrt{\check{\gamma}_h^{\text{VB}} \widehat{\delta}_h^{\text{VB}}}, \quad \widehat{b}_h = \pm \sqrt{\frac{\check{\gamma}_h^{\text{VB}}}{\widehat{\delta}_h^{\text{VB}}}}, \quad \widehat{\sigma}_{a_h}^2 = \frac{\sigma'^2 \widehat{\delta}_h^{\text{VB}}}{N\gamma_h}, \quad \widehat{\sigma}_{b_h}^2 = \frac{\sigma'^2}{N\gamma_h \widehat{\delta}_h^{\text{VB}}},
\end{aligned} \tag{14.25}$$

$$\text{where } \widehat{\delta}_h^{\text{VB}} \left(\equiv \frac{\widehat{a}_h}{\widehat{b}_h} \right) = \begin{cases} \frac{(\max(L,M)-\min(L,M))c_{a_h}}{\gamma_h} + O_p(1) & (\text{if } L \leq M), \\ \left(\frac{(\max(L,M)-\min(L,M))c_{b_h}}{\gamma_h} + O_p(1) \right)^{-1} & (\text{if } L > M), \end{cases} \quad (14.26)$$

and otherwise,

$$\widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = c_{a_h}^2 \left(1 - \frac{NL\widehat{\zeta}_h^{\text{VB}}}{\sigma'^2} \right), \quad \widehat{\sigma}_{b_h}^2 = c_{b_h}^2 \left(1 - \frac{NM\widehat{\zeta}_h^{\text{VB}}}{\sigma'^2} \right), \quad (14.27)$$

$$\text{where } \widehat{\zeta}_h^{\text{VB}} \left(\equiv \widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2 \right) = \begin{cases} \frac{\min(L,M)\sigma'^2}{NLM} + \Theta(N^{-2}), & (\text{if } L \neq M), \\ \frac{\min(L,M)\sigma'^2}{NLM} + \Theta(N^{-3/2}), & (\text{if } L = M). \end{cases} \quad (14.28)$$

Proof Noting that Corollary 6.8 gives the VB posterior for either \mathbf{V} or $\mathbf{V}^\top \in \mathbb{R}^{L \times M}$ that satisfies $L \leq M$, we have Eq. (14.25) with

$$\begin{aligned} \widehat{\delta}_h^{\text{VB}} &= \begin{cases} \frac{Nc_{a_h}}{\sigma'^2} \left(\gamma_h - \check{\gamma}_h^{\text{VB}} - \frac{L\sigma'^2}{N\gamma_h} \right) & (\text{if } L \leq M) \\ \left(\frac{Nc_{b_h}}{\sigma'^2} \left(\gamma_h - \check{\gamma}_h^{\text{VB}} - \frac{M\sigma'^2}{N\gamma_h} \right) \right)^{-1} & (\text{if } L > M) \end{cases} \\ &= \begin{cases} \frac{(\max(L,M)-\min(L,M))c_{a_h}}{\gamma_h} + O_p(1) & (\text{if } L \leq M), \\ \left(\frac{(\max(L,M)-\min(L,M))c_{b_h}}{\gamma_h} + O_p(1) \right)^{-1} & (\text{if } L > M), \end{cases} \end{aligned}$$

when $\gamma_h > \underline{\gamma}_h^{\text{VB}}$, and Eq. (14.27) with

$$\begin{aligned} \widehat{\zeta}_h^{\text{VB}} &= \frac{\sigma'^2}{2NLM} \left\{ L + M + \frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} - \sqrt{\left(L + M + \frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} \right)^2 - 4LM} \right\} \\ &= \frac{\sigma'^2}{2NLM} \left\{ L + M + \frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} \right. \\ &\quad \left. - \sqrt{(L+M)^2 + 2(L+M) \frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} + \left(\frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} \right)^2 - 4LM} \right\} \\ &= \frac{\sigma'^2}{2NLM} \left\{ L + M + \frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} \right. \\ &\quad \left. - \sqrt{(\max(L,M) - \min(L,M))^2 + 2(L+M) \frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} + \left(\frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} \right)^2} \right\} \end{aligned}$$

$$\begin{aligned}
& \left(\frac{\sigma'^2}{2NLM} \left\{ L + M + \frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} - (\max(L, M) - \min(L, M)) \right. \right. \\
& \quad \cdot \left(1 + \frac{L+M}{(\max(L, M) - \min(L, M))^2} \frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} \right) + O(N^{-2}) \Big\} \quad (\text{if } L \neq M) \\
& = \left\{ \frac{\sigma'^2}{2NLM} \left\{ L + M + \frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} \right. \right. \\
& \quad \left. \left. - \sqrt{2(L+M) \left(\frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} \right) + \left(\frac{\sigma'^2}{Nc_{a_h}^2 c_{b_h}^2} \right)^2} \right\} \quad (\text{if } L = M) \right. \\
& = \left\{ \frac{\sigma'^2}{2NLM} (2 \min(L, M) + \Theta(N^{-1})) \quad (\text{if } L \neq M) \right. \\
& \quad \left. \frac{\sigma'^2}{2NLM} (2 \min(L, M) + \Theta(N^{-1/2})) \quad (\text{if } L = M) \right. \\
& = \left\{ \frac{\min(L, M)\sigma'^2}{NLM} + \Theta(N^{-2}) \quad (\text{if } L \neq M), \right. \\
& \quad \left. \frac{\min(L, M)\sigma'^2}{NLM} + \Theta(N^{-3/2}) \quad (\text{if } L = M), \right.
\end{aligned}$$

when $\gamma_h \leq \underline{\gamma}_h^{\text{VB}}$. This completes the proof. \square

From Corollary 14.2, we can evaluate the orders of the optimal variational parameters in the asymptotic limit, which will be used when the VB free energy is analyzed.

Corollary 14.3 *The orders of the optimal variational parameters, given by Eq. (14.25) or Eq. (14.27), are as follows: if $\gamma_h > \underline{\gamma}_h^{\text{VB}} (= \Theta(N^{-1/2}))$,*

$$\begin{aligned}
& \widehat{a}_h = \Theta_p(1), \quad \widehat{b}_h = \Theta_p(\gamma_h), \quad \widehat{\sigma}_{a_h}^2 = \Theta_p(N^{-1}\gamma_h^{-2}), \quad \widehat{\sigma}_{b_h}^2 = \Theta_p(N^{-1}) \quad (\text{if } L < M), \\
& \widehat{a}_h = \Theta_p(\gamma_h^{1/2}), \quad \widehat{b}_h = \Theta_p(\gamma_h^{1/2}), \quad \widehat{\sigma}_{a_h}^2 = \Theta_p(N^{-1}\gamma_h^{-1}), \quad \widehat{\sigma}_{b_h}^2 = \Theta_p(N^{-1}\gamma_h^{-1}) \quad (\text{if } L = M), \\
& \widehat{a}_h = \Theta_p(\gamma_h), \quad \widehat{b}_h = \Theta_p(1), \quad \widehat{\sigma}_{a_h}^2 = \Theta_p(N^{-1}), \quad \widehat{\sigma}_{b_h}^2 = \Theta_p(N^{-1}\gamma_h^{-2}) \quad (\text{if } L > M),
\end{aligned} \tag{14.29}$$

and otherwise,

$$\begin{aligned}
& \widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = \Theta(1), \quad \widehat{\sigma}_{b_h}^2 = \Theta(N^{-1}) \quad (\text{if } L < M), \\
& \widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = \Theta(N^{-1/2}), \quad \widehat{\sigma}_{b_h}^2 = \Theta(N^{-1/2}) \quad (\text{if } L = M), \\
& \widehat{a}_h = 0, \quad \widehat{b}_h = 0, \quad \widehat{\sigma}_{a_h}^2 = \Theta(N^{-1}), \quad \widehat{\sigma}_{b_h}^2 = \Theta(1) \quad (\text{if } L > M).
\end{aligned} \tag{14.30}$$

Proof Eqs. (14.22) and (14.23) give

$$\underline{\gamma}_h^{\text{VB}} = \Theta(N^{-1/2}), \quad \check{\gamma}_h^{\text{VB}} = \Theta(\gamma_h),$$

and Eq. (14.26) gives

$$\widehat{\delta}_h^{\text{VB}} = \begin{cases} \Theta_p(\gamma_h^{-1}) & (\text{if } L < M), \\ \Theta_p(1) & (\text{if } L = M), \\ \Theta_p(\gamma_h) & (\text{if } L > M). \end{cases}$$

Substituting the preceding into Eq. (14.25) gives Eq. (14.29), and substituting Eq. (14.28) into Eq. (14.27) gives Eq. (14.30), which complete the proof. \square

Corollary 14.3 implies that the posterior probability mass does not necessarily converge to a single point, for example, $\widehat{\sigma}_{a_h}^2 = \Theta(1)$ if $\gamma_h < \underline{\gamma}_h^{\text{VB}}$ and $L < M$. This is typical behavior of singular models with *nonidentifiability*. On the other hand, the probability mass of the linear mapping $\mathbf{U} = \mathbf{B}\mathbf{A}^\top$ converges to a single point.

Corollary 14.4 *It holds that*

$$\left\langle \left\| \mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \right\|_{\text{Fro}}^2 \right\rangle_{r_A(A)r_B(\mathbf{B})} = O_p(N^{-1}).$$

Proof We have

$$\begin{aligned} \left\langle \left\| \mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \right\|_{\text{Fro}}^2 \right\rangle_{r_A(A)r_B(\mathbf{B})} &= \text{tr} \left\langle \left(\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \right)^\top \left(\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \right) \right\rangle_{r_A(A)r_B(\mathbf{B})} \\ &= \text{tr} \left\langle \mathbf{A}\mathbf{B}^\top \mathbf{B}\mathbf{A}^\top - 2\mathbf{A}\mathbf{B}^\top \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top + \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \right\rangle_{r_A(A)r_B(\mathbf{B})} \\ &= \text{tr} \left\langle \left(\mathbf{A}^\top \mathbf{A}\mathbf{B}^\top \mathbf{B} \right)_{r_A(A)r_B(\mathbf{B})} - \widehat{\mathbf{A}}^\top \widehat{\mathbf{A}}\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} \right\rangle \\ &= \text{tr} \left\langle \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M\widehat{\Sigma}_A \right) \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L\widehat{\Sigma}_B \right) - \widehat{\mathbf{A}}^\top \widehat{\mathbf{A}}\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} \right\rangle \\ &= \sum_{h=1}^H \left((\widehat{a}_h^2 + M\widehat{\sigma}_{a_h}^2) (\widehat{b}_h^2 + L\widehat{\sigma}_{b_h}^2) - \widehat{a}_h^2 \widehat{b}_h^2 \right) \\ &= \sum_{h=1}^H \left(L\widehat{a}_h^2 \widehat{\sigma}_{b_h}^2 + M\widehat{b}_h^2 \widehat{\sigma}_{a_h}^2 + LM\widehat{\sigma}_{a_h}^2 \widehat{\sigma}_{b_h}^2 \right). \end{aligned} \tag{14.31}$$

Corollary 14.3 guarantees that all terms in Eq. (14.31) are of the order of $\Theta_p(N^{-1})$ for any L, M , and $\{\gamma_h\}$, which completes the proof. \square

Now we derive an asymptotic form of the VB predictive distribution,

$$p(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{Y}) = \langle p(\mathbf{y}|\mathbf{x}, \mathbf{A}, \mathbf{B}) \rangle_{r_A(A)r_B(\mathbf{B})}. \tag{14.32}$$

From Corollary 14.4, we expect that the predictive distribution is not very far from the *plug-in* VB predictive distribution (see Section 1.1.3):

$$p(y|x, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \text{Gauss}_L(y; \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x}, \sigma'^2 \mathbf{I}_L). \quad (14.33)$$

Indeed, we will show in the next section that both predictive distributions (14.32) and (14.33) give the same generalization and training coefficients. This justifies the use of the *plug-in* VB predictive distribution, which is easy to compute from the optimal variational parameters.

By expanding the VB predictive distribution around the plug-in VB predictive distribution, we have the following theorem:

Theorem 14.5 *The VB predictive distribution (14.32) of the RRR model (14.2) and (14.3) can be written as*

$$p(y|x, X, Y) = \text{Gauss}_L(y; \boldsymbol{\Psi} \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x}, \sigma'^2 \boldsymbol{\Psi}) + O_p(N^{-3/2}) \quad (14.34)$$

for $\boldsymbol{\Psi} = \mathbf{I}_L + O_p(N^{-1})$.

Proof The VB predictive distribution can be written as follows:

$$\begin{aligned} p(y|x, X, Y) &= \langle p(y|x, \mathbf{A}, \mathbf{B}) \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\ &= p(y|x, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}) \left\langle \frac{p(y|x, \mathbf{A}, \mathbf{B})}{p(y|x, \widehat{\mathbf{A}}, \widehat{\mathbf{B}})} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\ &= p(y|x, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}) \left\langle \exp \left(-\frac{\|y - \mathbf{B}\mathbf{A}^\top \mathbf{x}\|^2 - \|y - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x}\|^2}{2\sigma'^2} \right) \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\ &= p(y|x, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}) \left\langle \exp \left(-\frac{(y - \mathbf{B}\mathbf{A}^\top \mathbf{x} + (y - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x}))^\top (y - \mathbf{B}\mathbf{A}^\top \mathbf{x} - (y - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x}))}{2\sigma'^2} \right) \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\ &= p(y|x, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}) \left\langle \exp \left(\frac{(y - (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top)\mathbf{x})^\top (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top)\mathbf{x}}{\sigma'^2} \right) \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}. \end{aligned} \quad (14.35)$$

Corollary 14.4 implies that the exponent in Eq. (14.35) is of the order of $N^{-1/2}$, i.e.,

$$\phi \equiv \frac{(y - (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top)\mathbf{x})^\top (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top)\mathbf{x}}{\sigma'^2} = O_p(N^{-1/2}). \quad (14.36)$$

By applying the Taylor expansion of the exponential function to Eq. (14.35), we obtain an asymptotic expansion of the predictive distribution around the plug-in predictive distribution:

$$p(y|x, X, Y) = p(y|x, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}) \left(1 + \langle \phi \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} + \frac{1}{2} \langle \phi^2 \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} + O_p(N^{-3/2}) \right).$$

Focusing on the dependence on the random variable \mathbf{y} , we can identify the function form of the predictive distribution as follows:

$$\begin{aligned}
p(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{Y}) &\propto \exp\left(-\frac{\|\mathbf{y} - \widehat{\mathbf{BA}}^\top \mathbf{x}\|^2}{2\sigma'^2}\right) \\
&\quad + \log\left(1 + \langle \phi \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} + \frac{1}{2} \langle \phi^2 \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} + O_p(N^{-3/2})\right) \\
&= \exp\left(-\frac{\|\mathbf{y} - \widehat{\mathbf{BA}}^\top \mathbf{x}\|^2}{2\sigma'^2} + \langle \phi \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} + \frac{1}{2} \langle \phi^2 \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}\right. \\
&\quad \left.- \frac{1}{2} \langle \phi \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}^2 + O_p(N^{-3/2})\right) \\
&\propto \exp\left(-\frac{\|\mathbf{y} - \widehat{\mathbf{BA}}^\top \mathbf{x}\|^2}{2\sigma'^2} + \frac{1}{2} \langle \phi^2 \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} + O_p(N^{-3/2})\right) \\
&\propto \exp\left(-\frac{\|\mathbf{y} - \widehat{\mathbf{BA}}^\top \mathbf{x}\|^2 - \mathbf{y}^\top \boldsymbol{\Psi}_1 \mathbf{y}}{2\sigma'^2} + O_p(N^{-3/2})\right) \\
&\propto \exp\left(-\frac{\|\mathbf{y}\|^2 - 2\mathbf{y}^\top \widehat{\mathbf{BA}}^\top \mathbf{x} - \mathbf{y}^\top \boldsymbol{\Psi}_1 \mathbf{y}}{2\sigma'^2} + O_p(N^{-3/2})\right) \\
&\propto \exp\left(-\frac{(\mathbf{y} - \boldsymbol{\Psi} \widehat{\mathbf{BA}}^\top \mathbf{x})^\top \boldsymbol{\Psi}^{-1} (\mathbf{y} - \boldsymbol{\Psi} \widehat{\mathbf{BA}}^\top \mathbf{x})}{2\sigma'^2} + O_p(N^{-3/2})\right), \tag{14.37}
\end{aligned}$$

where

$$\boldsymbol{\Psi} = (\mathbf{I}_L - \boldsymbol{\Psi}_1)^{-1}, \tag{14.38}$$

$$\boldsymbol{\Psi}_1 = \left\langle \frac{(\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\mathbf{A}^\top) \mathbf{x} \mathbf{x}^\top (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\mathbf{A}^\top)^\top}{\sigma'^2} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})}. \tag{14.39}$$

Here we used

$$\begin{aligned}
\langle \phi \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} &= \left\langle \frac{(\mathbf{y} - (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\mathbf{A}^\top)\mathbf{x})^\top (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\mathbf{A}^\top)\mathbf{x}}{\sigma'^2} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\
&= \left\langle \frac{\|(\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\mathbf{A}^\top)\mathbf{x}\|^2}{\sigma'^2} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\
&= \text{const.}, \\
\langle \phi^2 \rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} &= \left\langle \frac{(\mathbf{y} - (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\mathbf{A}^\top)\mathbf{x})^\top (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\mathbf{A}^\top) \mathbf{x} \mathbf{x}^\top (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\mathbf{A}^\top)^\top (\mathbf{y} - (\mathbf{B}\mathbf{A}^\top - \widehat{\mathbf{B}}\mathbf{A}^\top)\mathbf{x})}{\sigma'^4} \right\rangle_{r_A(\mathbf{A})r_B(\mathbf{B})} \\
&= \frac{\mathbf{y}^\top \boldsymbol{\Psi}_1 \mathbf{y}}{\sigma'^2} + O_p(N^{-3/2}).
\end{aligned}$$

Eq. (14.39) implies that $\boldsymbol{\Psi}_1$ is symmetric and $\boldsymbol{\Psi}_1 = O_p(N^{-1})$. Therefore, $\boldsymbol{\Psi}$, defined by Eq. (14.38), is symmetric, positive definite, and can be written

as $\boldsymbol{\Psi} = \mathbf{I}_L + O_p(N^{-1})$. The function form of Eq. (14.37) implies that the VB predictive distribution converges to the Gaussian distribution in the asymptotic limit, and we thus have

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{Y}) &= \frac{\exp\left(-\frac{(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})^\top \boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})}{2\sigma'^2} + O_p(N^{-3/2})\right)}{\int \exp\left(-\frac{(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})^\top \boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})}{2\sigma'^2} + O_p(N^{-3/2})\right) d\mathbf{y}} \\ &= \frac{\exp\left(-\frac{(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})^\top \boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})}{2\sigma'^2}\right)(1+O_p(N^{-3/2}))}{\int \exp\left(-\frac{(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})^\top \boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})}{2\sigma'^2}\right)(1+O_p(N^{-3/2})) d\mathbf{y}} \\ &= \frac{1}{(2\pi\sigma'^2)^{L/2} \det(\boldsymbol{\Psi})^{1/2}} \exp\left(-\frac{(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})^\top \boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \mathbf{x})}{2\sigma'^2}\right) + O_p(N^{-3/2}), \end{aligned}$$

which completes the proof. \square

14.2 Generalization Properties

Let us analyze generalization properties of VB learning based on the posterior distribution and the predictive distribution, derived in Section 14.1.2.

14.2.1 Assumption on True Distribution

We assume that the true distribution can be expressed by the model distribution with the *true* parameter \mathbf{A}^* and \mathbf{B}^* with their rank H^* :

$$\begin{aligned} q(\mathbf{y}|\mathbf{x}) &= \text{Gauss}_L\left(\mathbf{y}, \mathbf{B}^*\mathbf{A}^{*\top}\mathbf{x}, \sigma'^2\mathbf{I}_L\right) \\ &= (2\pi\sigma'^2)^{-L/2} \exp\left(-\frac{\|\mathbf{y} - \mathbf{B}^*\mathbf{A}^{*\top}\mathbf{x}\|^2}{2\sigma'^2}\right). \end{aligned} \quad (14.40)$$

Let

$$\mathbf{U}^* \equiv \mathbf{B}^*\mathbf{A}^{*\top} = \sum_{h=1}^{\min(L,M)} \gamma_h^* \boldsymbol{\omega}_{b_h}^* \boldsymbol{\omega}_{a_h}^{*\top} \quad (14.41)$$

be the SVD of the true linear mapping $\mathbf{B}^*\mathbf{A}^{*\top}$, where γ_h^* (≥ 0) is the h th largest singular value, and $\boldsymbol{\omega}_{a_h}^*$ and $\boldsymbol{\omega}_{b_h}^*$ are the associated right and left singular vectors. The assumption that the true linear mapping has rank H^* amounts to

$$\gamma_h^* = \begin{cases} \Theta(1) & \text{for } h = 1, \dots, H^*, \\ 0 & \text{for } h = H^* + 1, \dots, \min(L, M). \end{cases} \quad (14.42)$$

14.2.2 Consistency of VB Estimator

Since the training samples are drawn from the true distribution (14.40), the central limit theorem (Theorem 13.1) guarantees the following:

$$\begin{aligned} V \left(\equiv \frac{1}{N} \sum_{n=1}^N \mathbf{y}^{(n)} \mathbf{x}^{(n)\top} \right) &= \frac{1}{N} \sum_{n=1}^N (\mathbf{B}^* \mathbf{A}^{*\top} \mathbf{x}^{(n)} + \boldsymbol{\varepsilon}^{(n)}) \mathbf{x}^{(n)\top} \\ &= \mathbf{B}^* \mathbf{A}^{*\top} + O_p(N^{-1/2}), \end{aligned} \quad (14.43)$$

$$\begin{aligned} \langle \mathbf{x} \mathbf{x}^\top \rangle_{q(\mathbf{x})} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)} \mathbf{x}^{(n)\top} + O_p(N^{-1/2}) \\ &= \mathbf{I}_M + O_p(N^{-1/2}). \end{aligned} \quad (14.44)$$

Here we used the assumption (14.7) that the input is prewhitened. Eq. (14.43) is consistent with Eq. (14.9), which implies that the distribution of \mathbf{V} is given by

$$q(\mathbf{V}) = \text{MGauss}_{L,M} \left(\mathbf{V}; \mathbf{B}^* \mathbf{A}^{*\top}, \frac{\sigma'^2}{N} \mathbf{I}_L \otimes \mathbf{I}_M \right), \quad (14.45)$$

and therefore

$$\langle \mathbf{V} \rangle_{q(X,Y)} = \langle \mathbf{V} \rangle_{q(V)} = \mathbf{B}^* \mathbf{A}^{*\top}, \quad (14.46)$$

and for each (l, m) ,

$$\left\langle \left\| \mathbf{V}_{l,m} - (\mathbf{B}^* \mathbf{A}^{*\top})_{l,m} \right\|_{\text{Fro}}^2 \right\rangle_{q(X,Y)} = \frac{\sigma'^2}{N}. \quad (14.47)$$

Eq. (14.43) implies that

$$\gamma_h = \gamma_h^* + O_p(N^{-1/2}), \quad (14.48)$$

where γ_h is the h th largest singular value of \mathbf{V} (see Eq. (14.15)). Eq. (14.45) also implies that, for any h ,

$$\sum_{h': \gamma_{h'}^* = \gamma_h^*} \left\langle \gamma_{h'} \omega_{b_{h'}} \omega_{a_{h'}}^\top \right\rangle_{q(X,Y)} = \sum_{h': \gamma_{h'}^* = \gamma_h^*} \gamma_{h'}^* \omega_{b_{h'}}^* \omega_{a_{h'}}^{*\top}, \quad (14.49)$$

$$\sum_{h': \gamma_{h'}^* = \gamma_h^*} \gamma_{h'} \omega_{b_{h'}} \omega_{a_{h'}}^\top = \sum_{h': \gamma_{h'}^* = \gamma_h^*} \gamma_{h'}^* \omega_{b_{h'}}^* \omega_{a_{h'}}^{*\top} + O_p(N^{-1/2}), \quad (14.50)$$

where $\sum_{h': \gamma_{h'}^* = \gamma_h^*}$ denotes the sum over all h' such that $\gamma_{h'}^* = \gamma_h^*$. Eq. (14.50) implies that for any nonzero and nondegenerate singular component h (i.e., $\gamma_h^* > 0$ and $\gamma_h^* \neq \gamma_{h'}^* \forall h' \neq h$), it holds that

$$\begin{aligned}\omega_{a_h} &= \omega_{a_h}^* + O_p(N^{-1/2}), \\ \omega_{b_h} &= \omega_{b_h}^* + O_p(N^{-1/2}).\end{aligned}$$

Eq. (14.9) implies that the ML estimator is given by

$$\left(\widehat{\mathbf{BA}}^\top\right)^{\text{ML}} = \sum_{h=1}^H \gamma_h \omega_{b_h} \omega_{a_h}^\top. \quad (14.51)$$

Therefore, Eq. (14.43) guarantees the convergence of the ML estimator to the true linear mapping $\mathbf{B}^* \mathbf{A}^{*\top}$ when $H \geq H^*$.

Lemma 14.6 (*Consistency of ML estimator in RRR*) It holds that

$$\left(\widehat{\mathbf{BA}}^\top\right)^{\text{ML}} - \mathbf{B}^* \mathbf{A}^{*\top} = \begin{cases} \Theta(1) & \text{if } H < H^*, \\ O_p(N^{-1/2}) & \text{if } H \geq H^*. \end{cases}$$

We can also show the convergence of the VB estimator:

Lemma 14.7 (*Consistency of VB estimator in RRR*) It holds that

$$\widehat{\mathbf{BA}}^\top - \mathbf{B}^* \mathbf{A}^{*\top} = \begin{cases} \Theta(1) & \text{if } H < H^*, \\ O_p(N^{-1/2}) & \text{if } H \geq H^*. \end{cases}$$

Proof The case where $H < H^*$ is trivial because the rank H matrix $\widehat{\mathbf{BA}}^\top$ can never converge to the rank H^* matrix $\mathbf{B}^* \mathbf{A}^{*\top}$. Assume that $H \geq H^*$. Theorem 14.1 implies that, when $\gamma_h > \underline{\gamma}_h^{\text{VB}} (= \Theta(N^{-1/2}))$,

$$\widehat{\gamma}_h^{\text{VB}} = \check{\gamma}_h^{\text{VB}} = \gamma_h \left(1 - \frac{\max(L, M)\sigma'^2}{N\underline{\gamma}_h^2}\right) + O_p(N^{-1}) = \gamma_h + O_p(N^{-1/2}),$$

and otherwise

$$\widehat{\gamma}_h^{\text{VB}} = 0.$$

Since $\gamma_h = O_p(N^{-1/2})$ for $h = H^* + 1, \dots, \min(L, M)$, the preceding two equations lead to

$$\begin{aligned}\widehat{\mathbf{BA}}^\top &= \sum_{h=1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top = \sum_{h=1}^{\min(L, M)} \gamma_h \omega_{b_h} \omega_{a_h}^\top + O_p(N^{-1/2}) = \mathbf{V} + O_p(N^{-1/2}).\end{aligned} \quad (14.52)$$

Substituting Eq. (14.43) into Eq. (14.52) completes the proof. \square

14.2.3 Generalization Error

Now we analyze the asymptotic behavior of the generalization error. We first show the asymptotic equivalence between the VB predictive distribution,

given by Theorem 14.5, and the plug-in VB predictive distribution (14.33)—both give the same leading term of the generalization error with $O_p(N^{-3/2})$ difference. To this end, we use the following lemma:

Lemma 14.8 *For any three sets of Gaussian parameters (μ^*, Σ^*) , $(\widehat{\mu}, \widehat{\Sigma})$, $(\acute{\mu}, \acute{\Sigma})$ such that*

$$\widehat{\mu} = \mu^* + O_p(N^{-1/2}), \quad \widehat{\Sigma} = \Sigma^* + O_p(N^{-1/2}), \quad (14.53)$$

$$\acute{\mu} = \widehat{\mu} + O_p(N^{-1}), \quad \acute{\Sigma} = \widehat{\Sigma} + O_p(N^{-1}), \quad (14.54)$$

it holds that

$$\left\langle \log \frac{\text{Gauss}_L(y; \acute{\mu}, \acute{\Sigma}) + O_p(N^{-3/2})}{\text{Gauss}_L(y; \widehat{\mu}, \widehat{\Sigma})} \right\rangle_{\text{Gauss}_L(y; \mu^*, \Sigma^*)} = O_p(N^{-3/2}). \quad (14.55)$$

Proof The (twice of the) left-hand side of Eq. (14.55) can be written as

$$\begin{aligned} \psi_1 &\equiv 2 \left\langle \log \frac{\text{Gauss}_L(y; \acute{\mu}, \acute{\Sigma}) + O_p(N^{-3/2})}{\text{Gauss}_L(y; \widehat{\mu}, \widehat{\Sigma})} \right\rangle_{\text{Gauss}_L(y; \mu^*, \Sigma^*)} \\ &= \left\langle \log \frac{\det(\widehat{\Sigma})}{\det(\acute{\Sigma})} - (y - \acute{\mu})^\top \acute{\Sigma}^{-1} (y - \acute{\mu}) + (y - \widehat{\mu})^\top \widehat{\Sigma}^{-1} (y - \widehat{\mu}) \right\rangle_{\text{Gauss}_L(y; \mu^*, \Sigma^*)} \\ &\quad + O_p(N^{-3/2}) \\ &= -\log \det(\Sigma^* \widehat{\Sigma}^{-1} \Sigma^{*-1} \acute{\Sigma}) \\ &\quad - \left\langle (y - \mu^* - (\acute{\mu} - \mu^*))^\top \acute{\Sigma}^{-1} (y - \mu^* - (\acute{\mu} - \mu^*)) \right\rangle_{\text{Gauss}_L(y; \mu^*, \Sigma^*)} \\ &\quad + \left\langle (y - \mu^* - (\widehat{\mu} - \mu^*))^\top \widehat{\Sigma}^{-1} (y - \mu^* - (\widehat{\mu} - \mu^*)) \right\rangle_{\text{Gauss}_L(y; \mu^*, \Sigma^*)} + O_p(N^{-3/2}) \\ &= \text{tr} \left(\log(\Sigma^* \acute{\Sigma}^{-1}) - \log(\Sigma^* \widehat{\Sigma}^{-1}) \right) - \text{tr} \left(\Sigma^* \acute{\Sigma}^{-1} \right) - (\acute{\mu} - \mu^*)^\top \acute{\Sigma}^{-1} (\acute{\mu} - \mu^*) \\ &\quad + \text{tr} \left(\Sigma^* \widehat{\Sigma}^{-1} \right) + (\widehat{\mu} - \mu^*)^\top \widehat{\Sigma}^{-1} (\widehat{\mu} - \mu^*) + O_p(N^{-3/2}). \end{aligned}$$

By using Eqs. (14.53) and (14.54) and the Taylor expansion of the logarithmic function, we have

$$\begin{aligned} \psi_1 &= \text{tr} \left(\left(\Sigma^* \acute{\Sigma}^{-1} - \mathbf{I}_L \right) - \frac{\left(\Sigma^* \acute{\Sigma}^{-1} - \mathbf{I}_L \right)^\top \left(\Sigma^* \acute{\Sigma}^{-1} - \mathbf{I}_L \right)}{2} \right. \\ &\quad \left. - \left(\Sigma^* \widehat{\Sigma}^{-1} - \mathbf{I}_L \right) + \frac{\left(\Sigma^* \widehat{\Sigma}^{-1} - \mathbf{I}_L \right)^\top \left(\Sigma^* \widehat{\Sigma}^{-1} - \mathbf{I}_L \right)}{2} \right) \\ &\quad - \text{tr} \left(\Sigma^* \acute{\Sigma}^{-1} \right) - (\acute{\mu} - \mu^*)^\top \acute{\Sigma}^{-1} (\acute{\mu} - \mu^*) \end{aligned}$$

$$\begin{aligned}
& + \text{tr} \left(\Sigma^* \widehat{\Sigma}^{-1} \right) + (\widehat{\mu} - \mu^*)^\top \widehat{\Sigma}^{-1} (\widehat{\mu} - \mu^*) + O_p(N^{-3/2}) \\
& = O_p(N^{-3/2}),
\end{aligned}$$

which completes the proof. \square

Given a test input x , Lemma 14.8 can be applied to the true distribution (14.40), the plug-in VB predictive distribution (14.33), and the predictive distribution (14.34) when $H \geq H^*$, where

$$\begin{aligned}
\mu^* &= B^* A^{*\top} x, & \Sigma^* &= \sigma'^2 I_L, \\
\widehat{\mu} &= \widehat{B} \widehat{A}^\top x = \mu^* + O_p(N^{-1/2}), & \widehat{\Sigma} &= \sigma'^2 I_L = \Sigma^*, \\
\acute{\mu} &= \Psi \widehat{B} \widehat{A}^\top x = \widehat{\mu} + O_p(N^{-1}), & \acute{\Sigma} &= \sigma'^2 \Psi = \widehat{\Sigma} + O_p(N^{-1}),
\end{aligned}$$

for $\Psi = I_L + O_p(N^{-1})$. Here, Lemma 14.7 was used in the equation for $\widehat{\mu}$. Thus, we have the following corollary:

Corollary 14.9 *When $H \geq H^*$, it holds that*

$$\left\langle \log \frac{p(y|x, X, Y)}{p(y|x, \widehat{A}, \widehat{B})} \right\rangle_{q(y|x)} = O_p(N^{-3/2}),$$

and therefore the difference between the generalization error (13.30) of the VB predictive distribution (14.34) and the generalization error of the plug-in VB predictive distribution (14.33) is of the order of $N^{-3/2}$, i.e.,

$$\begin{aligned}
\text{GE}(\mathcal{D}) &= \left\langle \log \frac{q(y|x)}{p(y|x, X, Y)} \right\rangle_{q(y|x)q(x)} \\
&= \left\langle \log \frac{q(y|x)}{p(y|x, \widehat{A}, \widehat{B})} \right\rangle_{q(y|x)q(x)} + O_p(N^{-3/2}).
\end{aligned}$$

Corollary 14.9 leads to the following theorem:

Theorem 14.10 *The generalization error of the RRR model is written as*

$$\text{GE}(\mathcal{D}) = \begin{cases} \Theta(1) & \text{if } H < H^*, \\ \frac{\|\widehat{B} \widehat{A}^\top - B^* A^{*\top}\|_{\text{Fro}}^2}{2\sigma'^2} + O_p(N^{-3/2}) & \text{if } H \geq H^*. \end{cases} \quad (14.56)$$

Proof When $H < H^*$, Theorem 14.5 implies that

$$\begin{aligned}
\text{GE}(\mathcal{D}) &= \left\langle \log \frac{q(y|x)}{p(y|x, \widehat{A}, \widehat{B})} \right\rangle_{q(y|x)q(x)} + O_p(N^{-1}) \\
&= \frac{\|\widehat{B} \widehat{A}^\top - B^* A^{*\top}\|_{\text{Fro}}^2}{2\sigma'^2} + O_p(N^{-1}).
\end{aligned}$$

With Lemma 14.7, we have $\text{GE}(\mathcal{D}) = \Theta(1)$. When $H \geq H^*$, we have

$$\begin{aligned}\text{GE}(\mathcal{D}) &= \left\langle \log \frac{q(y|\mathbf{x})}{p(y|\mathbf{x}, \widehat{\mathbf{A}}, \widehat{\mathbf{B}})} \right\rangle_{q(y|\mathbf{x})q(\mathbf{x})} + O_p(N^{-3/2}) \\ &= \left\langle -\frac{\|y - \mathbf{B}^* \mathbf{A}^{*\top} \mathbf{x}\|^2 - \|y - \widehat{\mathbf{B}} \widehat{\mathbf{A}}^{*\top} \mathbf{x}\|^2}{2\sigma'^2} \right\rangle_{q(y|\mathbf{x})q(\mathbf{x})} + O_p(N^{-3/2}) \\ &= \left\langle -\frac{\|y - \mathbf{B}^* \mathbf{A}^{*\top} \mathbf{x}\|^2 - \|y - \mathbf{B}^* \mathbf{A}^{*\top} \mathbf{x} - (\widehat{\mathbf{B}} \widehat{\mathbf{A}}^{*\top} - \mathbf{B}^* \mathbf{A}^{*\top}) \mathbf{x}\|^2}{2\sigma'^2} \right\rangle_{q(y|\mathbf{x})q(\mathbf{x})} + O_p(N^{-3/2}) \\ &= \left\langle \frac{\|(\widehat{\mathbf{B}} \widehat{\mathbf{A}}^{*\top} - \mathbf{B}^* \mathbf{A}^{*\top}) \mathbf{x}\|^2}{2\sigma'^2} \right\rangle_{q(\mathbf{x})} + O_p(N^{-3/2}) \\ &= \left\langle \frac{\text{tr} \left\{ (\widehat{\mathbf{B}} \widehat{\mathbf{A}}^{*\top} - \mathbf{B}^* \mathbf{A}^{*\top}) \mathbf{x} \mathbf{x}^\top (\widehat{\mathbf{B}} \widehat{\mathbf{A}}^{*\top} - \mathbf{B}^* \mathbf{A}^{*\top})^\top \right\}}{2\sigma'^2} \right\rangle_{q(\mathbf{x})} + O_p(N^{-3/2}).\end{aligned}$$

By using Eq. (14.44) and Lemma 14.7, we obtain Eq. (14.56), which completes the proof. \square

Next we compute the average generalization error (13.13) over the distribution of training samples. As Theorem 14.10 states, the generalization error never converges to zero if $H < H^*$, since a rank H^* matrix cannot be well approximated by a rank H matrix. Accordingly, we hereafter focus on the case where $H \geq H^*$. By $\text{Wishart}_D(V, \nu)$ we denote the D -dimensional Wishart distribution with scale matrix V and degree of freedom ν . Then we have the following theorem:

Theorem 14.11 *The average generalization error of the RRR model for $H \geq H^*$ is asymptotically expanded as*

$$\overline{\text{GE}}(N) = \langle \text{GE}(\mathcal{D}) \rangle_{q(\mathcal{D})} = \lambda^{\text{VB}} N^{-1} + O(N^{-3/2}),$$

where the generalization coefficient is given by

$$\begin{aligned}2\lambda^{\text{VB}} &= (H^*(L+M) - H^{*2}) \\ &\quad + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > \max(L, M)) \left(1 - \frac{\max(L, M)}{\gamma_h'^2}\right)^2 \gamma_h'^2 \right\rangle_{q(\mathbf{W})}. \quad (14.57)\end{aligned}$$

Here $\gamma_h'^2$ is the h th largest eigenvalue of a random matrix $\mathbf{W} \in \mathbb{S}_+^{\min(L, M)}$ subject to $\text{Wishart}_{\min(L, M)-H^*}(\mathbf{I}_{\min(L, M)-H^*}, \max(L, M) - H^*)$, and $\theta(\cdot)$ is the indicator function such that $\theta(\text{condition}) = 1$ if the condition is true and $\theta(\text{condition}) = 0$ otherwise.

Proof Theorem 14.1 and Eqs. (14.42) and (14.48) imply that

$$\widehat{\gamma}_h^{\text{VB}} = \begin{cases} \gamma_h + O_p(N^{-1}) = O_p(1) & \text{for } h = 1, \dots, H^*, \\ \max\left(0, \gamma_h \left(1 - \frac{\max(L, M)\sigma'^2}{N\gamma_h^2}\right)\right) + O_p(N^{-1}) = O_p(N^{-1/2}) & \text{for } h = H^* + 1, \dots, H. \end{cases} \quad (14.58)$$

Therefore, we have

$$\begin{aligned} \left\| \widehat{\mathbf{BA}}^\top - \mathbf{B}^* \mathbf{A}^{*\top} \right\|_{\text{Fro}}^2 &= \left\| \sum_{h=1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top - \sum_{h=1}^{\min(L, M)} \gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top} \right\|_{\text{Fro}}^2 \\ &= \left\| \sum_{h=1}^{H^*} \left(\gamma_h \omega_{b_h} \omega_{a_h}^\top - \gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top} \right) + \sum_{h=H^*+1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top + O_p(N^{-1}) \right\|_{\text{Fro}}^2 \\ &= \left\| \sum_{h=1}^{H^*} \left(\gamma_h \omega_{b_h} \omega_{a_h}^\top - \gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top} \right) + \sum_{h=H^*+1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top \right\|_{\text{Fro}}^2 + O_p(N^{-3/2}) \\ &= \left\| \mathbf{V} - \mathbf{B}^* \mathbf{A}^* + \sum_{h=H^*+1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top - \sum_{h=H^*+1}^{\min(L, M)} \gamma_h \omega_{b_h} \omega_{a_h}^\top \right\|_{\text{Fro}}^2 \\ &\quad + O_p(N^{-3/2}). \end{aligned}$$

Here, in order to get the third equation, we used the fact that the first two terms in the norm in the second equation are of the order of $O_p(N^{-1/2})$. The expectation over the distribution of training samples is given by

$$\begin{aligned} &\left\langle \left\| \widehat{\mathbf{BA}}^\top - \mathbf{B}^* \mathbf{A}^{*\top} \right\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\ &= \left\langle \left\| \mathbf{V} - \mathbf{B}^* \mathbf{A}^* + \sum_{h=H^*+1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top - \sum_{h=H^*+1}^{\min(L, M)} \gamma_h \omega_{b_h} \omega_{a_h}^\top \right\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\ &\quad + O(N^{-3/2}) \\ &= \left\langle \|\mathbf{V} - \mathbf{B}^* \mathbf{A}^*\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\ &\quad + 2 \left\langle (\mathbf{V} - \mathbf{B}^* \mathbf{A}^*)^\top \left(\sum_{h=H^*+1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top - \sum_{h=H^*+1}^{\min(L, M)} \gamma_h \omega_{b_h} \omega_{a_h}^\top \right) \right\rangle_{q(\mathcal{D})} \\ &\quad + \left\langle \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}})^2 - 2 \sum_{h=H^*+1}^H \gamma_h \widehat{\gamma}_h^{\text{VB}} + \sum_{h=H^*+1}^{\min(L, M)} \gamma_h^2 \right\rangle_{q(\mathcal{D})} \\ &\quad + O(N^{-3/2}) \\ &= \left\langle \|\mathbf{V} - \mathbf{B}^* \mathbf{A}^*\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} + 2 \left\langle \sum_{h=H^*+1}^H \gamma_h \widehat{\gamma}_h^{\text{VB}} - \sum_{h=H^*+1}^{\min(L, M)} \gamma_h^2 \right\rangle_{q(\mathcal{D})} \\ &\quad + \left\langle \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}})^2 - 2 \sum_{h=H^*+1}^H \gamma_h \widehat{\gamma}_h^{\text{VB}} + \sum_{h=H^*+1}^{\min(L, M)} \gamma_h^2 \right\rangle_{q(\mathcal{D})} + O(N^{-3/2}) \\ &= \left\langle \|\mathbf{V} - \mathbf{B}^* \mathbf{A}^*\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} - \left\langle \sum_{h=H^*+1}^{\min(L, M)} \gamma_h^2 \right\rangle_{q(\mathcal{D})} \\ &\quad + \left\langle \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}})^2 \right\rangle_{q(\mathcal{D})} + O(N^{-3/2}). \quad (14.59) \end{aligned}$$

Here we used Eq. (14.49) and the orthonormality of the singular vectors.

Eq. (14.45) implies that the first term in Eq. (14.59) is equal to

$$\left\langle \|\mathbf{V} - \mathbf{B}^* \mathbf{A}^*\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} = LM \frac{\sigma'^2}{N}. \quad (14.60)$$

The redundant components $\{\gamma_h \omega_{b_h} \omega_{a_h}^\top\}_{h=H^*+1}^{\min(L,M)}$ are zero-mean (see Eq. (14.49)) Gaussian matrices capturing the Gaussian noise in the orthogonal space to the necessary components $\{\gamma_h \omega_{b_h} \omega_{a_h}^\top\}_{h=1}^{H^*}$. Therefore, the distribution of the corresponding singular values $\{\gamma_h\}_{h=H^*+1}^{\min(L,M)}$ coincides with the distribution of the singular values of $\mathbf{V}' \in \mathbb{R}^{(\min(L,M)-H^*) \times (\max(L,M)-H^*)}$ subject to

$$q(\mathbf{V}') = \text{MGauss}_{\min(L,M)-H^*, \max(L,M)-H^*} \left(\mathbf{V}'; \mathbf{0}_{\min(L,M)-H^*, \max(L,M)-H^*}, \frac{\sigma'^2}{N} \mathbf{I}_{\min(L,M)-H^*} \otimes \mathbf{I}_{\max(L,M)-H^*} \right). \quad (14.61)$$

This leads to

$$\left\langle \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^2 \right\rangle_{q(\mathcal{D})} = (L - H^*)(M - H^*) \frac{\sigma'^2}{N}. \quad (14.62)$$

Let $\{\gamma'_h\}_{h=1}^{\min(L,M)-H^*}$ be the singular values of $\frac{\sqrt{N}}{\sigma'} \mathbf{V}'$. Then, $\{\gamma'_h\}_{h=1}^{\min(L,M)-H^*}$ are the eigenvalues of $\mathbf{W} = \frac{N}{\sigma'^2} \mathbf{V}' \mathbf{V}'^\top$, which is subject to Wishart _{$\min(L,M)-H^*$} ($\mathbf{I}_{\min(L,M)-H^*}, \max(L, M) - H^*$). By substituting Eqs. (14.60), (14.62), and (14.58) into Eq. (14.59), we have

$$\begin{aligned} & \left\langle \|\widehat{\mathbf{BA}}^\top - \mathbf{B}^* \mathbf{A}^*\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\ &= \frac{\sigma'^2}{N} \{LM - (L - H^*)(M - H^*)\} \\ &+ \left\langle \sum_{h=H^*+1}^H \left\{ \max \left(0, \gamma_h \left(1 - \frac{\max(L,M)\sigma'^2}{N\gamma_h^2} \right) \right) \right\}^2 \right\rangle_{q(\mathcal{D})} + O(N^{-3/2}) \\ &= \frac{\sigma'^2}{N} \left\{ (H^*(L + M) - H^{*2}) + \left\langle \sum_{h=1}^{H-H^*} \left\{ \max \left(0, 1 - \frac{\max(L,M)}{\gamma_h^2} \right) \right\}^2 \gamma_h^2 \right\rangle_{q(\mathbf{W})} \right\} \\ &+ O(N^{-3/2}). \end{aligned}$$

Substituting the preceding into Eq. (14.56) completes the proof. \square

The first and the second terms in Eq. (14.57) correspond to the contribution from the necessary components $h = 1, \dots, H^*$ and the contribution from the redundant components $h = H^* + 1, \dots, H$, respectively. If we focus on the parameter space of the first H^* components, i.e., $\{\mathbf{a}_h, \mathbf{b}_h\}_{h=1}^{H^*}$, the true linear mapping $\{\mathbf{a}_h^*, \mathbf{b}_h^*\}_{h=1}^{H^*}$ lies at an (essentially) nonsingular point (after removing

the trivial H^{*2} redundancy). Therefore, as the regular learning theory states, the contribution from the necessary components is equal to the (essential) degree of freedom (see Eq. (14.4)) of the RRR model for $H = H^*$. On the other hand, the regular learning theory cannot be applied to the redundant components $\{\mathbf{a}_h, \mathbf{b}_h\}_{h=H^*+1}^H$ since the true parameter is on the *singularities* $\{\mathbf{a}_h = \mathbf{0}\} \cup \{\mathbf{b}_h = \mathbf{0}\}$, making the second term different from the degree of freedom of the redundant parameters.

Assuming that L and M are large, we can approximate the second term in Eq. (14.57) by using the *random matrix theory* (see Section 8.4.1). Consider the large-scale limit when L, M, H, H^* go to infinity with the same ratio, so that

$$\alpha = \frac{\min(L, M) - H^*}{\max(L, M) - H^*}, \quad (14.63)$$

$$\beta = \frac{H - H^*}{\min(L, M) - H^*}, \quad (14.64)$$

$$\kappa = \frac{\max(L, M)}{\max(L, M) - H^*} \quad (14.65)$$

are constant. Then Marčenko–Pastur law (Proposition 8.11) states that the empirical distribution of the eigenvalues $\{y_1, \dots, y_{\min(L,M)-H^*}\}$ of the random matrix $\frac{NVV^\top}{(\max(L,M)-H^*)\sigma^2} \sim \text{Wishart}_{\min(L,M)-H^*}(\mathbf{I}_{\min(L,M)-H^*}, 1)$ almost surely converges to

$$p(y) \rightarrow p^{\text{MP}}(y) \equiv \frac{\sqrt{(y - \underline{y})(\bar{y} - y)}}{2\pi\alpha y} \theta(\underline{y} < y < \bar{y}), \quad (14.66)$$

$$\text{where } \bar{y} = (1 + \sqrt{\alpha})^2, \quad \underline{y} = (1 - \sqrt{\alpha})^2. \quad (14.67)$$

Let

$$(2\pi\alpha)^{-1} J'_s(u) = \int_u^\infty y^s p(y) dy \quad (14.68)$$

be the s th order (incomplete) moment of the Marčenko–Pastur distribution (14.66) with the lower bound u of the integration range. Then, the second term of Eq. (14.57) can be written as

$$\begin{aligned} & \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > \max(L, M)) \left(1 - \frac{\max(L, M)}{\gamma_h'^2}\right)^2 \gamma_h'^2 \right\rangle_{q(W)} \\ & \rightarrow (\min(L, M) - H^*)(\max(L, M) - H^*) \int_{u_\beta}^\infty \theta(y > \kappa) \left(1 - \frac{\kappa}{y}\right)^2 y p(y) dy \\ & = (\min(L, M) - H^*)(\max(L, M) - H^*) \int_{\max(\kappa, u_\beta)}^\infty \left(y - 2\kappa + \kappa^2 y^{-1}\right) p(y) dy \\ & = \frac{(\min(L, M) - H^*)(\max(L, M) - H^*)}{2\pi\alpha} \left(J'_1(\hat{u}) - 2\kappa J'_0(\hat{u}) + \kappa^2 J'_{-1}(\hat{u}) \right), \end{aligned}$$

where u_β is the β -percentile point of $p(y)$, i.e.,

$$\beta = \int_{u_\beta}^{\infty} p(y) dy = (2\pi\alpha)^{-1} J'_0(u_\beta), \quad (14.69)$$

and

$$\hat{u} = \max(\kappa, u_\beta). \quad (14.70)$$

Using the transformation $z = (\underline{y} - (\underline{y} + \bar{y})/2) / (2\sqrt{\alpha})$, we can derive analytic forms of the moments (14.68) and thus obtain the following theorem:

Theorem 14.12 *The VB generalization coefficient of the RRR model in the large-scale limit is given by*

$$2\lambda^{\text{VB}} \rightarrow (H^*(L+M) - H^{*2}) + \frac{(\min(L, M) - H^*)(\max(L, M) - H^*)}{2\pi\alpha} \left\{ J_1(\hat{z}) - 2\kappa J_0(\hat{z}) + \kappa^2 J_{-1}(\hat{z}) \right\}, \quad (14.71)$$

where

$$J_1(z) = 2\alpha(-z\sqrt{1-z^2} + \cos^{-1} z),$$

$$J_0(z) = -2\sqrt{\alpha}\sqrt{1-z^2} + (1+\alpha)\cos^{-1} z - (1-\alpha)\cos^{-1} \frac{\sqrt{\alpha}(1+\alpha)z + 2\alpha}{2\alpha z + \sqrt{\alpha}(1+\alpha)},$$

$$J_{-1}(z) = \begin{cases} 2\sqrt{\alpha} \frac{\sqrt{1-z^2}}{2\sqrt{\alpha}z+1+\alpha} - \cos^{-1} z + \frac{1+\alpha}{1-\alpha} \cos^{-1} \frac{\sqrt{\alpha}(1+\alpha)z + 2\alpha}{2\alpha z + \sqrt{\alpha}(1+\alpha)} & (0 < \alpha < 1), \\ 2\sqrt{\frac{1-z}{1+z}} - \cos^{-1} z & (\alpha = 1), \end{cases}$$

and $\hat{z} = \max((\kappa - (1 + \alpha))/2\sqrt{\alpha}, J_0^{-1}(2\pi\alpha\beta))$. Here $J_s^{-1}(\cdot)$ denotes the inverse function of $J_s(z)$.

Theorem 14.12 allows us to compare the generalization error of VB learning with those of ML (MAP) learning and Bayesian learning in Section 14.2.6.

14.2.4 Training Error

The training error can be analyzed in a similar way to the generalization error. We first prove the following lemma:

Lemma 14.13 *Let $\bar{\mathbf{U}} \in \mathbb{R}^{L,M}$ and $\bar{\Sigma} \in \mathbb{S}_+^L$ be the ML estimators of the linear regression model $\mathbf{y} = \mathbf{U}\mathbf{x} + \boldsymbol{\varepsilon}$ with Gaussian noise $\boldsymbol{\varepsilon} \sim \text{Gauss}(\mathbf{0}, \Sigma)$. For any two sets of parameters $(\bar{\mathbf{U}}, \bar{\Sigma}), (\hat{\mathbf{U}}, \hat{\Sigma})$ such that*

$$\hat{\mathbf{U}} = \bar{\mathbf{U}} + O_p(N^{-1/2}), \quad \hat{\Sigma} = \bar{\Sigma} + O_p(N^{-1/2}), \quad (14.72)$$

$$\hat{\mathbf{U}} = \hat{\mathbf{U}} + O_p(N^{-1}), \quad \hat{\Sigma} = \hat{\Sigma} + O_p(N^{-1}), \quad (14.73)$$

it holds that

$$\frac{1}{N} \sum_{n=1}^N \log \frac{\text{Gauss}_L(\mathbf{y}^{(n)}; \hat{\mathbf{U}}\mathbf{x}^{(n)}, \hat{\Sigma}) + O_p(N^{-3/2})}{\text{Gauss}_L(\mathbf{y}^{(n)}; \hat{\mathbf{U}}\mathbf{x}^{(n)}, \hat{\Sigma})} = O_p(N^{-3/2}). \quad (14.74)$$

Proof The (twice of the) left-hand side of Eq. (14.74) can be written as

$$\begin{aligned} \psi_2 &\equiv \frac{2}{N} \sum_{n=1}^N \log \frac{\text{Gauss}_L(\mathbf{y}^{(n)}; \hat{\mathbf{U}}\mathbf{x}^{(n)}, \hat{\Sigma}) + O_p(N^{-3/2})}{\text{Gauss}_L(\mathbf{y}^{(n)}; \hat{\mathbf{U}}\mathbf{x}^{(n)}, \hat{\Sigma})} \\ &= \log \frac{\det(\hat{\Sigma})}{\det(\hat{\Sigma})} + \frac{1}{N} \sum_{n=1}^N \left(-(\mathbf{y}^{(n)} - \hat{\mathbf{U}}\mathbf{x}^{(n)})^\top \hat{\Sigma}^{-1} (\mathbf{y}^{(n)} - \hat{\mathbf{U}}\mathbf{x}^{(n)}) \right. \\ &\quad \left. + (\mathbf{y}^{(n)} - \hat{\mathbf{U}}\mathbf{x}^{(n)})^\top \hat{\Sigma}^{-1} (\mathbf{y}^{(n)} - \hat{\mathbf{U}}\mathbf{x}^{(n)}) \right) + O_p(N^{-3/2}) \\ &= -\log \det(\bar{\Sigma}\hat{\Sigma}^{-1}\bar{\Sigma}^{-1}\hat{\Sigma}) + O_p(N^{-3/2}) \\ &\quad - \frac{1}{N} \sum_{n=1}^N \left(\mathbf{y}^{(n)} - \bar{\mathbf{U}}\mathbf{x}^{(n)} - (\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right)^\top \hat{\Sigma}^{-1} \left(\mathbf{y}^{(n)} - \bar{\mathbf{U}}\mathbf{x}^{(n)} - (\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right) \\ &\quad + \frac{1}{N} \sum_{n=1}^N \left(\mathbf{y}^{(n)} - \bar{\mathbf{U}}\mathbf{x}^{(n)} - (\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right)^\top \hat{\Sigma}^{-1} \left(\mathbf{y}^{(n)} - \bar{\mathbf{U}}\mathbf{x}^{(n)} - (\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right) \\ &= \text{tr} \left(\log(\bar{\Sigma}\hat{\Sigma}^{-1}) - \log(\bar{\Sigma}\hat{\Sigma}^{-1}) \right) + O_p(N^{-3/2}) \\ &\quad - \text{tr}(\bar{\Sigma}\hat{\Sigma}^{-1}) - \frac{1}{N} \sum_{n=1}^N \left((\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right)^\top \hat{\Sigma}^{-1} \left((\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right) \\ &\quad + \text{tr}(\bar{\Sigma}\hat{\Sigma}^{-1}) + \frac{1}{N} \sum_{n=1}^N \left((\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right)^\top \hat{\Sigma}^{-1} \left((\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right). \end{aligned}$$

By using Eqs. (14.72) and (14.73) and the Taylor expansion of the logarithmic function, we have

$$\begin{aligned} \psi_2 &= \text{tr} \left((\bar{\Sigma}\hat{\Sigma}^{-1} - \mathbf{I}_L) - \frac{(\bar{\Sigma}\hat{\Sigma}^{-1} - \mathbf{I}_L)^\top (\bar{\Sigma}\hat{\Sigma}^{-1} - \mathbf{I}_L)}{2} - (\bar{\Sigma}\hat{\Sigma}^{-1} - \mathbf{I}_L) + \frac{(\bar{\Sigma}\hat{\Sigma}^{-1} - \mathbf{I}_L)^\top (\bar{\Sigma}\hat{\Sigma}^{-1} - \mathbf{I}_L)}{2} \right) \\ &\quad - \text{tr}(\bar{\Sigma}\hat{\Sigma}^{-1}) - \frac{1}{N} \sum_{n=1}^N \left((\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right)^\top \hat{\Sigma}^{-1} \left((\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right) \\ &\quad + \text{tr}(\bar{\Sigma}\hat{\Sigma}^{-1}) + \frac{1}{N} \sum_{n=1}^N \left((\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right)^\top \hat{\Sigma}^{-1} \left((\hat{\mathbf{U}} - \bar{\mathbf{U}})\mathbf{x}^{(n)} \right) + O_p(N^{-3/2}) \\ &= O_p(N^{-3/2}), \end{aligned}$$

which completes the proof. \square

When $H \geq H^*$, Lemma 14.13 can be applied to the plug-in VB predictive distribution (14.33) and the VB predictive distribution (14.34), where

$$\begin{aligned}\bar{\mathbf{U}} &= \mathbf{V}, & \bar{\Sigma} &= \frac{\sigma'^2}{N} \sum_{n=1}^N (\mathbf{y}^{(n)} - \mathbf{V}\mathbf{x}^{(n)}) (\mathbf{y}^{(n)} - \mathbf{V}\mathbf{x}^{(n)})^\top \\ & & &= \sigma'^2 \mathbf{I}_L + O_p(N^{-1/2}), \\ \widehat{\mathbf{U}} &= \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top = \bar{\mathbf{U}} + O_p(N^{-1/2}), & \widehat{\Sigma} &= \sigma'^2 \mathbf{I}_L = \bar{\Sigma} + O_p(N^{-1/2}), \\ \dot{\mathbf{U}} &= \boldsymbol{\Psi}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top = \widehat{\mathbf{U}} + O_p(N^{-1}), & \dot{\Sigma} &= \sigma'^2 \boldsymbol{\Psi} = \widehat{\Sigma} + O_p(N^{-1}),\end{aligned}$$

for $\boldsymbol{\Psi} = \mathbf{I}_L + O_p(N^{-1})$. Here Eq. (14.43) and Lemma 14.7 were used in the equation for $\widehat{\mathbf{U}}$. Thus, we have the following corollary:

Corollary 14.14 *When $H \geq H^*$, it holds that*

$$\frac{1}{N} \sum_{n=1}^N \log \frac{p(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, \mathbf{X}, \mathbf{Y})}{p(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, \widehat{\mathbf{A}}, \widehat{\mathbf{B}})} = O_p(N^{-3/2}),$$

and therefore the difference between the training error (13.31) of the VB predictive distribution (14.34) and the training error of the plug-in VB predictive distribution (14.33) is of the order of $N^{-3/2}$, i.e.,

$$\begin{aligned}\text{TE}(\mathcal{D}) &= \frac{1}{N} \sum_{n=1}^N \log \frac{q(\mathbf{y}^{(n)} | \mathbf{x}^{(n)})}{p(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, \mathbf{X}, \mathbf{Y})} \\ &= \frac{1}{N} \sum_{n=1}^N \log \frac{q(\mathbf{y}^{(n)} | \mathbf{x}^{(n)})}{p(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, \widehat{\mathbf{A}}, \widehat{\mathbf{B}})} + O_p(N^{-3/2}).\end{aligned}$$

Corollary 14.14 leads to the following theorem:

Theorem 14.15 *The training error of the RRR model is written as*

$$\text{TE}(\mathcal{D}) = \begin{cases} \Theta(1) & \text{if } H < H^*, \\ \frac{\|V - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top\|_{\text{Fro}}^2 - \|V - \mathbf{B}^*\mathbf{A}^{*\top}\|_{\text{Fro}}^2}{2\sigma'^2} + O_p(N^{-3/2}) & \text{if } H \geq H^*. \end{cases} \quad (14.75)$$

Proof When $H < H^*$, Theorem 14.5 implies that

$$\begin{aligned}\text{TE}(\mathcal{D}) &= \frac{1}{N} \sum_{n=1}^N \log \frac{q(\mathbf{y}^{(n)} | \mathbf{x}^{(n)})}{p(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, \widehat{\mathbf{A}}, \widehat{\mathbf{B}})} + O_p(N^{-1}) \\ &= \frac{\|V - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top\|_{\text{Fro}}^2 - \|V - \mathbf{B}^*\mathbf{A}^{*\top}\|_{\text{Fro}}^2}{2\sigma'^2} + O_p(N^{-1}).\end{aligned}$$

With Lemma 14.7, we have $\text{TE}(\mathcal{D}) = \Theta(1)$. When $H \geq H^*$, we have

$$\begin{aligned}\text{TE}(\mathcal{D}) &= -\frac{1}{N} \sum_{n=1}^N \frac{\|y^{(n)} - \mathbf{B}^* \mathbf{A}^{*\top} \mathbf{x}^{(n)}\|^2 - \|y^{(n)} - \widehat{\mathbf{BA}}^\top \mathbf{x}^{(n)}\|^2}{2\sigma^2} + O_p(N^{-3/2}) \\ &= -\frac{1}{N} \sum_{n=1}^N \frac{\|y^{(n)} - \mathbf{V} \mathbf{x}^{(n)} - (\mathbf{B}^* \mathbf{A}^{*\top} - \mathbf{V}) \mathbf{x}^{(n)}\|^2 - \|y^{(n)} - \mathbf{V} \mathbf{x}^{(n)} - (\widehat{\mathbf{BA}}^\top - \mathbf{V}) \mathbf{x}^{(n)}\|^2}{2\sigma^2} + O_p(N^{-3/2}) \\ &= -\frac{1}{N} \sum_{n=1}^N \frac{\|(\mathbf{B}^* \mathbf{A}^{*\top} - \mathbf{V}) \mathbf{x}^{(n)}\|^2 - \|(\widehat{\mathbf{BA}}^\top - \mathbf{V}) \mathbf{x}^{(n)}\|^2}{2\sigma^2} + O_p(N^{-3/2}) \\ &= -\frac{1}{N} \sum_{n=1}^N \frac{\text{tr}\left((\mathbf{B}^* \mathbf{A}^{*\top} - \mathbf{V}) \mathbf{x}^{(n)} \mathbf{x}^{(n)\top} (\mathbf{B}^* \mathbf{A}^{*\top} - \mathbf{V})^\top - (\widehat{\mathbf{BA}}^\top - \mathbf{V}) \mathbf{x}^{(n)} \mathbf{x}^{(n)\top} (\widehat{\mathbf{BA}}^\top - \mathbf{V})^\top\right)}{2\sigma^2} \\ &\quad + O_p(N^{-3/2}).\end{aligned}$$

By using the prewhitening condition (14.7) and Lemma 14.7, we obtain Eq. (14.75), which completes the proof. \square

Now we can derive an asymptotic form of the average training error:

Theorem 14.16 *The average training error of the RRR model for $H \geq H^*$ is asymptotically expanded as*

$$\overline{\text{TE}}(N) = \langle \text{TE}(\mathcal{D}) \rangle_{q(\mathcal{D})} = \nu^{\text{VB}} N^{-1} + O(N^{-3/2}),$$

where the training coefficient is given by

$$\begin{aligned}2\nu^{\text{VB}} &= -(H^*(L+M) - H^{*2}) \\ &- \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > \max(L, M)) \left(1 - \frac{\max(L, M)}{\gamma_h'^2}\right) \left(1 + \frac{\max(L, M)}{\gamma_h'^2}\right) \gamma_h'^2 \right\rangle_{q(W)}. \tag{14.76}\end{aligned}$$

Here $\gamma_h'^2$ is the h th largest eigenvalue of a random matrix $\mathbf{W} \in \mathbb{S}_+^{\min(L, M)}$ subject to $\text{Wishart}_{\min(L, M)-H^*}(\mathbf{I}_{\min(L, M)-H^*}, \max(L, M) - H^*)$.

Proof From Eq. (14.58), we have

$$\begin{aligned}\|\widehat{\mathbf{BA}}^\top - \mathbf{V}\|_{\text{Fro}}^2 &= \left\| \sum_{h=1}^H \widehat{\gamma}_h^{\text{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top - \sum_{h=1}^{\min(L, M)} \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top \right\|_{\text{Fro}}^2 \\ &= \left\| \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}} - \gamma_h) \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top - \sum_{h=H+1}^{\min(L, M)} \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top + O_p(N^{-1}) \right\|_{\text{Fro}}^2 \\ &= \left\| \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}} - \gamma_h) \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top - \sum_{h=H+1}^{\min(L, M)} \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top \right\|_{\text{Fro}}^2 + O_p(N^{-3/2}) \\ &= \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}} - \gamma_h)^2 + \sum_{h=H+1}^{\min(L, M)} \gamma_h^2 + O_p(N^{-3/2}) \\ &= \sum_{h=H^*+1}^H \left(\max\left(0, \gamma_h \left(1 - \frac{\max(L, M)\sigma^2}{N\gamma_h^2}\right)\right) - \gamma_h \right)^2 \\ &\quad + \sum_{h=H+1}^{\min(L, M)} \gamma_h^2 + O_p(N^{-3/2})\end{aligned}$$

$$\begin{aligned}
&= - \sum_{h=H^*+1}^H \theta\left(\gamma_h^2 > \frac{\max(L,M)\sigma'^2}{N}\right) \cdot \left(\gamma_h - \frac{\max(L,M)\sigma'^2}{N\gamma_h}\right) \left(\gamma_h + \frac{\max(L,M)\sigma'^2}{N\gamma_h}\right) \\
&\quad + \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^2 + O_p(N^{-3/2}).
\end{aligned} \tag{14.77}$$

By using Eqs. (14.60), (14.62) and (14.77), we have

$$\begin{aligned}
&\left\langle \left\| V - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top \right\|_{\text{Fro}}^2 - \left\| V - \mathbf{B}^* \mathbf{A}^{*\top} \right\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\
&= -\frac{\sigma'^2}{N} \left\{ (H^*(L+M) - H^{*2}) \right. \\
&\quad \left. + \sum_{h=H^*+1}^H \theta\left(\gamma_h^2 > \frac{\max(L,M)\sigma'^2}{N}\right) \cdot \left(\gamma_h - \frac{\max(L,M)\sigma'^2}{N\gamma_h}\right) \left(\gamma_h + \frac{\max(L,M)\sigma'^2}{N\gamma_h}\right) \right\} \\
&\quad + O_p(N^{-3/2}).
\end{aligned}$$

Thus, by introducing the singular values $\{\gamma'_h\}_{h=1}^{\min(L,M)-H^*}$ of $\frac{\sqrt{N}}{\sigma'} V'$, where V' is a random matrix subject to Eq. (14.61), and using Theorem 14.15, we obtain Eq. (14.76), which completes the proof. \square

Finally, we apply the Marčenko–Pastur law (Proposition 8.11) for evaluating the second term in Eq. (14.76). In the large-scale limit when L, M, H, H^* go to infinity with the same ratio, so that Eqs. (14.63) through (14.65) are constant, we have

$$\begin{aligned}
&\left\langle \sum_{h=1}^{H-H^*} \theta\left(\gamma'_h > \max(L, M)\right) \left(1 - \frac{\max(L, M)}{\gamma'_h^2}\right) \left(1 + \frac{\max(L, M)}{\gamma'_h^2}\right) \gamma'_h \right\rangle_{q(W)} \\
&\rightarrow (\min(L, M) - H^*)(\max(L, M) - H^*) \int_{u_\beta}^{\infty} \theta(y > \kappa) \left(1 - \frac{\kappa}{y}\right) \left(1 + \frac{\kappa}{y}\right) y p(y) dy \\
&= (\min(L, M) - H^*)(\max(L, M) - H^*) \int_{\max(\kappa, u_\beta)}^{\infty} (y - \kappa^2 y^{-1}) p(y) dy \\
&= \frac{(\min(L, M) - H^*)(\max(L, M) - H^*)}{2\pi\alpha} \left(J'_1(\hat{u}) - \kappa^2 J'_{-1}(\hat{u})\right),
\end{aligned}$$

where $J'_s(u)$, β , and \hat{u} are defined in Eqs. (14.68), (14.69), and (14.70), respectively. Thus, the transformation $z = (\underline{y} - (\underline{y} + \bar{y})/2)/(2\sqrt{\alpha})$ gives the following theorem:

Theorem 14.17 *The VB training coefficient of the RRR model in the large scale limit is given by*

$$\begin{aligned}
2\nu^{\text{VB}} &\rightarrow -(H^*(L+M) - H^{*2}) \\
&\quad - \frac{(\min(L, M) - H^*)(\max(L, M) - H^*)}{2\pi\alpha} \left\{ J_1(\hat{z}) - \kappa^2 J_{-1}(\hat{z}) \right\},
\end{aligned} \tag{14.78}$$

where $J_1(z)$, $J_{-1}(z)$, and \hat{z} are defined in Theorem 14.12.

14.2.5 Free Energy

The VB free energy can be analyzed relatively easily based on the orders of the variational parameters, given by Corollary 14.3:

Theorem 14.18 *The relative VB free energy (13.150) of the RRR model for $H \geq H^*$ is asymptotically expanded as*

$$\tilde{F}^{\text{VB}}(\mathcal{D}) = F^{\text{VB}}(\mathbf{Y}|\mathbf{X}) - NS_N(\mathbf{Y}|\mathbf{X}) = \lambda'^{\text{VB}} \log N + O_p(1), \quad (14.79)$$

where the free energy coefficient is given by

$$2\lambda'^{\text{VB}} = H^*(L + M) + (H - H^*) \min(L, M). \quad (14.80)$$

Proof The VB free energy for the RRR model is given by Eq. (14.14), and the empirical entropy is given by

$$\begin{aligned} 2S_N(\mathbf{Y}|\mathbf{X}) &= -\frac{2}{N} \sum_{n=1}^N \log q(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}) \\ &= L \log(2\pi\sigma'^2) + \frac{\sum_{n=1}^N \|\mathbf{y}^{(n)} - \mathbf{B}^* \mathbf{A}^{*\top} \mathbf{x}^{(n)}\|^2}{N\sigma'^2} \\ &= L \log(2\pi\sigma'^2) + \frac{\frac{1}{N} \sum_{n=1}^N \|\mathbf{y}^{(n)}\|^2 - 2\text{tr}(\mathbf{V}^\top \mathbf{B}^* \mathbf{A}^{*\top}) + \|\mathbf{B}^* \mathbf{A}^{*\top}\|_{\text{Fro}}^2}{\sigma'^2} \\ &= L \log(2\pi\sigma'^2) + \frac{\frac{1}{N} \sum_{n=1}^N \|\mathbf{y}^{(n)}\|^2 + \|\mathbf{V} - \mathbf{B}^* \mathbf{A}^{*\top}\|_{\text{Fro}}^2 - \|\mathbf{V}\|_{\text{Fro}}^2}{\sigma'^2}. \end{aligned}$$

Therefore, the relative VB free energy (14.79) is given as

$$\begin{aligned} 2\tilde{F}^{\text{VB}}(\mathcal{D}) &= N \cdot \frac{\|\mathbf{V} - \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top\|_{\text{Fro}}^2 - \|\mathbf{V} - \mathbf{B}^* \mathbf{A}^{*\top}\|_{\text{Fro}}^2}{\sigma'^2} + M \log \frac{\det(C_A)}{\det(\widehat{\Sigma}_A)} + L \log \frac{\det(C_B)}{\det(\widehat{\Sigma}_B)} \\ &\quad - (L + M)H + \text{tr} \left\{ \mathbf{C}_A^{-1} \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) + \mathbf{C}_B^{-1} \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B \right) \right. \\ &\quad \left. + N\sigma'^{-2} \left(-\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} + M \widehat{\Sigma}_A \right) \left(\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}} + L \widehat{\Sigma}_B \right) \right) \right\}. \end{aligned} \quad (14.81)$$

Eqs. (14.52) and (14.43) imply that the first term in Eq. (14.81) is $O_p(1)$. Corollary 14.3 with Eqs. (14.42) and (14.48) implies that, for $h = 1, \dots, H^*$,

$$\widehat{a}_h = \Theta_p(1), \quad \widehat{b}_h = \Theta_p(1), \quad \widehat{\sigma}_{a_h}^2 = \Theta_p(N^{-1}), \quad \widehat{\sigma}_{b_h}^2 = \Theta_p(N^{-1}),$$

and, for $h = H^* + 1, \dots, H$,

$$\begin{aligned} \widehat{a}_h &= O_p(1), \quad \widehat{b}_h = O_p(N^{-1/2}), \quad \widehat{\sigma}_{a_h}^2 = \Theta_p(1), \quad \widehat{\sigma}_{b_h}^2 = \Theta_p(N^{-1}), \quad (\text{if } L < M), \\ \widehat{a}_h &= O_p(N^{-1/4}), \quad \widehat{b}_h = O_p(N^{-1/4}), \quad \widehat{\sigma}_{a_h}^2 = \Theta_p(N^{-1/2}), \quad \widehat{\sigma}_{b_h}^2 = \Theta_p(N^{-1/2}), \quad (\text{if } L = M), \\ \widehat{a}_h &= O_p(N^{-1/2}), \quad \widehat{b}_h = O_p(1), \quad \widehat{\sigma}_{a_h}^2 = \Theta_p(N^{-1}), \quad \widehat{\sigma}_{b_h}^2 = \Theta_p(1), \quad (\text{if } L > M). \end{aligned}$$

These results imply that the most terms in Eq. (14.81) are $O_p(1)$, and we thus have

$$\begin{aligned} 2\bar{F}^{\text{VB}}(\mathcal{D}) &= M \log \frac{\det(\mathbf{C}_A)}{\det(\widehat{\Sigma}_A)} + L \log \frac{\det(\mathbf{C}_B)}{\det(\widehat{\Sigma}_B)} + O_p(1) \\ &= M \log \prod_{h=1}^H \widehat{\sigma}_{a_h}^{-2} + L \log \prod_{h=1}^H \widehat{\sigma}_{b_h}^{-2} + O_p(1) \\ &= \{H^*(L+M) + (H-H^*) \min(L, M)\} \log N + O_p(1), \end{aligned}$$

which completes the proof. \square

Clearly from the proof, the first term and the second term in Eq. (14.80) correspond to the contribution from the necessary components, $h = 1, \dots, H^*$, and the contribution from the redundant components, $h = H^*, \dots, H$, respectively. A remark is that the contribution from the necessary components contains the trivial redundancy, i.e., it is $H^*(L+M)$ instead of $H^*(L+M) - H^{*2}$. This is because the independence between \mathbf{A} and \mathbf{B} prevents the VB posterior distribution from extending along the trivial redundancy.

14.2.6 Comparison with Other Learning Algorithms

Theorems 14.12, 14.17, and 14.18 allow us to compute the generalization, the training, and the free energy coefficients of VB learning. We can now compare those properties with those of ML learning and Bayesian learning, which have been clarified for the RRR model. Note that MAP learning with a smooth and finite prior (e.g., the Gaussian prior (14.3)) with fixed hyperparameters is asymptotically equivalent to ML learning, and has the same generalization and training coefficients.

ML Learning

The generalization error of ML learning in the RRR model was analyzed (Fukumizu, 1999), based on the Marčenko–Pastur law (Proposition 8.11). Let γ'_h be the h th largest eigenvalue of a random matrix $\mathbf{W} \in \mathbb{S}_+^{\min(L,M)}$ subject to Wishart $_{\min(L,M)-H^*}(\mathbf{I}_{\min(L,M)-H^*}, \max(L, M) - H^*)$.

Theorem 14.19 (Fukumizu, 1999) *The average ML generalization error of the RRR model for $H \geq H^*$ is asymptotically expanded as*

$$\overline{\text{GE}}^{\text{ML}}(N) = \lambda^{\text{ML}} N^{-1} + O(N^{-3/2}),$$

where the generalization coefficient is given by

$$2\lambda^{\text{ML}} = (H^*(L+M) - H^{*2}) + \left\langle \sum_{h=1}^{H-H^*} \gamma'_h^2 \right\rangle_{q(\mathbf{W})}. \quad (14.82)$$

Theorem 14.20 (Fukumizu, 1999) *The ML generalization coefficient of the RRR model in the large-scale limit is given by*

$$2\lambda^{\text{ML}} \rightarrow (H^*(L + M) - H^{*2}) + \frac{(\min(L, M) - H^*)(\max(L, M) - H^*)}{2\pi\alpha} J_1(\hat{z}), \quad (14.83)$$

where $J_1(\cdot)$ and \hat{z} are defined in Theorem 14.12.

Actually, Theorems 14.11 and 14.12 were derived by extending Theorems 14.19 and 14.20 to VB learning. We can derive Theorems 14.19 and 14.20 in the same way as VB learning by replacing the VB estimator (14.58) with the ML estimator $\widehat{\gamma}_h^{\text{ML}} = \gamma_h$.

The training error can be similarly analyzed.

Theorem 14.21 *The average ML training error of the RRR model for $H \geq H^*$ is asymptotically expanded as*

$$\overline{\text{TE}}^{\text{ML}}(N) = \nu^{\text{ML}} N^{-1} + O(N^{-3/2}),$$

where the training coefficient is given by

$$2\nu^{\text{ML}} = -(H^*(L + M) - H^{*2}) - \left\langle \sum_{h=1}^{H-H^*} \gamma_h'^2 \right\rangle_{q(W)}. \quad (14.84)$$

Theorem 14.22 *The ML training coefficient of the RRR model in the large-scale limit is given by*

$$2\nu^{\text{ML}} \rightarrow -(H^*(L + M) - H^{*2}) - \frac{(\min(L, M) - H^*)(\max(L, M) - H^*)}{2\pi\alpha} J_1(\hat{z}), \quad (14.85)$$

where $J_1(\cdot)$ and \hat{z} are defined in Theorem 14.12.

A note is that Theorems 14.19 and 14.21 imply that the generalization coefficient and the training coefficient are antisymmetric in ML learning, i.e., $\lambda^{\text{ML}} = -\nu^{\text{ML}}$, while they are not antisymmetric in VB learning, i.e., $\lambda^{\text{VB}} \neq -\nu^{\text{VB}}$ (see Theorems 14.11 and 14.16).

Bayesian Learning

The Bayes free energy in the RRR model was clarified based on the singular learning theory (see Section 13.5.4).

Theorem 14.23 (Aoyagi and Watanabe, 2005) *The relative Bayes free energy (13.32) in the RRR model is asymptotically expanded as*

$$\begin{aligned}\widetilde{F}^{\text{Bayes}}(\mathcal{D}) &= F^{\text{Bayes}}(Y|X) - NS_N(Y|X) \\ &= \lambda'^{\text{Bayes}} \log N - (m-1) \log \log N + O_p(1),\end{aligned}$$

where the free energy coefficient, as well as the coefficient of the second leading term, is given as follows:

(i) When $L + H^* \leq M + H$, $M + H^* \leq L + H$, and $H^* + H \leq L + M$:

(a) If $L + M + H + H^*$ is even, then $m = 1$ and

$$2\lambda'^{\text{Bayes}} = \frac{-(H^* + H)^2 - (L - M)^2 + 2(H^* + H)(L + M)}{4}.$$

(b) If $L + M + H + H^*$ is odd, then $m = 2$ and

$$2\lambda'^{\text{Bayes}} = \frac{-(H^* + H)^2 - (L - M)^2 + 2(H^* + H)(L + M) + 1}{4}.$$

(ii) When $M + H < L + H^*$, then $m = 1$ and

$$2\lambda'^{\text{Bayes}} = HM - HH^* + LH^*.$$

(iii) When $L + H < M + H^*$, then $m = 1$ and

$$2\lambda'^{\text{Bayes}} = HL - HH^* + MH^*.$$

(iv) When $L + M < H + H^*$, then $m = 1$ and

$$2\lambda'^{\text{Bayes}} = LM.$$

Theorem 14.23 immediately informs us of the asymptotic behavior of the Bayes generalization error, based on Corollary 13.14.

Theorem 14.24 (Aoyagi and Watanabe, 2005) *The Bayes generalization error of the RRR model for $H \geq H^*$ is asymptotically expanded as*

$$\overline{\text{GE}}^{\text{Bayes}}(N) = \lambda^{\text{Bayes}} N^{-1} - (m-1)(N \log N)^{-1} + o((N \log N)^{-1}),$$

where $\lambda^{\text{Bayes}} = \lambda'^{\text{Bayes}}$ and m are given in Theorem 14.23.

Unfortunately, the Bayes training error has not been clarified yet.

Numerical Comparison

Let us visually compare the theoretically clarified generalization properties. Figures 14.1 through 14.4 show the generalization coefficients and the training coefficients of the RRR model under the following settings:

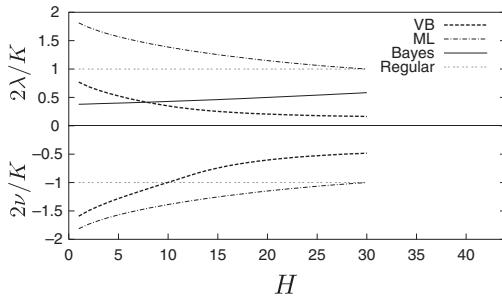


Figure 14.1 The generalization coefficients (in the positive vertical region) and the training coefficients (in the negative vertical region) of VB learning, ML learning, and Bayesian learning in the RRR model with $\max(L, M) = 50$, $\min(L, M) = 30$, $H = 1, \dots, 30$, and $H^* = 0$.

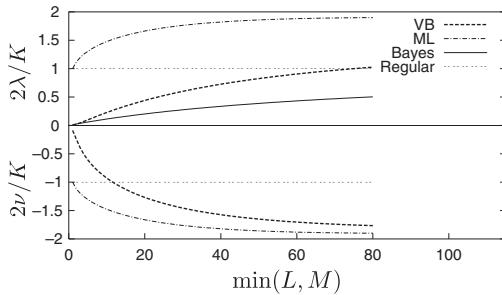


Figure 14.2 The generalization coefficients and the training coefficients ($\max(L, M) = 80$, $\min(L, M) = 1, \dots, 80$, $H = 1$, and $H^* = 0$).

- (i) $\max(L, M) = 50$, $\min(L, M) = 30$, $H = 1, \dots, 30$ (horizontal axis), $H^* = 0$,
- (ii) $\max(L, M) = 80$, $\min(L, M) = 1, \dots, 80$ (horizontal axis), $H = 1$, $H^* = 0$,
- (iii) $L = M = 80$, $H = 1, \dots, 80$ (horizontal axis), $H^* = 0$,
- (iv) $\max(L, M) = 50$, $\min(L, M) = 30$, $H = 20$, $H^* = 1, \dots, 20$ (horizontal axis).

The vertical axis indicates the coefficient normalized by the half of the *essential* parameter dimension D , given by Eq. (14.4). The curves in the positive vertical region correspond to the generalization coefficients of VB learning, ML learning, and Bayesian learning, while the curves in the negative vertical region correspond to the training coefficients. As a guide, we depicted the lines $2\lambda/D = 1$ and $2\nu/D = -1$, which correspond to the generalization and the training coefficients (by ML learning and Bayesian learning) of the regular

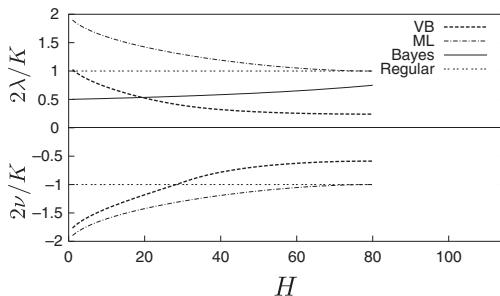


Figure 14.3 The generalization coefficients and the training coefficients ($L = M = 80$, $H = 1, \dots, 80$, and $H^* = 0$).

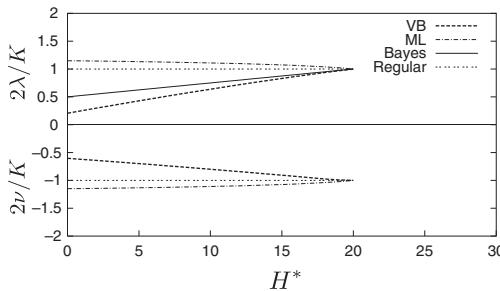


Figure 14.4 The generalization coefficients and the training coefficients ($\max(L, M) = 50$, $\min(L, M) = 30$, $H = 20$, and $H^* = 1, \dots, 20$).

models with the same parameter dimensionality. The curves for ML learning and VB learning were computed under the large-scale approximation, i.e., by using Theorems 14.12, 14.17, 14.20, and 14.22.¹

We see in Figures 14.1 through 14.4 that VB learning generally provides comparable generalization performance to Bayesian learning. However, significant differences are also observed. For example, we see in Figure 14.1 that VB learning provides much worse generalization performance than Bayesian learning when $H \ll \min(L, M)$, and much better performance when $H \sim \min(L, M)$.

Another finding is that, in Figures 14.1 and 14.3, the VB generalization coefficient depends on H similarly to the ML generalization coefficient. Moreover, we see that, when $\min(L, M) = 80$ in Figure 14.2 and when $H = 1$ in Figure 14.3, the VB generalization coefficient slightly exceeds the line

¹ We confirmed that numerical computation with Theorems 14.11, 14.16, 14.19, and 14.21 gives visually indistinguishable results.

$2\lambda/D = 1$ —the VB generalization coefficient per parameter dimension can be larger than that in the regular models, which never happens for the Bayes generalization coefficient (see Eq. (13.125)).

Finally, Figure 14.4 shows that, for this particular RRR model with $\max(L, M) = 50$, $\min(L, M) = 30$, and $H = 20$, VB learning always gives smaller generalization error than Bayesian learning in the asymptotic limit, regardless of the true rank H^* . This might be seen contradictory with the proven optimality of Bayesian learning—Bayesian learning is never dominated by any other method (see Appendix D for the optimality of Bayesian learning and Appendix A for the definition of the term “domination”). We further discuss this issue by considering *subtle true singular values* in Section 14.2.7.

Next we compare the VB free energy with the Bayes free energy, by using Theorems 14.18 and 14.23. Figures 14.5 through 14.8 show the free energy

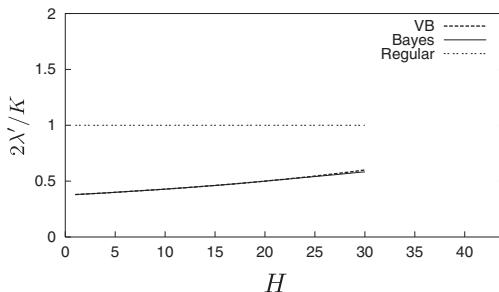


Figure 14.5 Free energy coefficients ($\max(L, M) = 50$, $\min(L, M) = 30$, $H = 1, \dots, 30$, and $H^* = 0$). The VB and the Bayes free energy coefficients are almost overlapped.

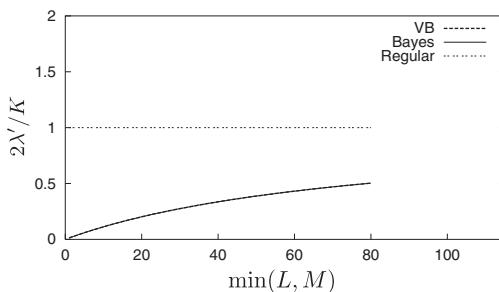


Figure 14.6 Free energy coefficients ($\max(L, M) = 80$, $\min(L, M) = 1, \dots, 80$, $H = 1$, and $H^* = 0$). The VB and the Bayes free energy coefficients are almost overlapped.

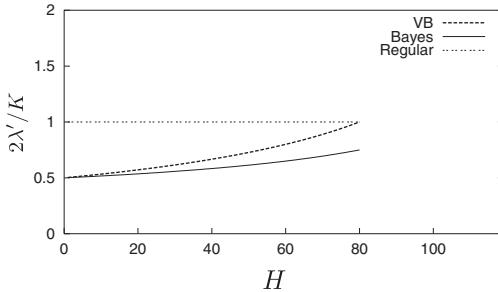


Figure 14.7 Free energy coefficients ($L = M = 80$, $H = 1, \dots, 80$, and $H^* = 0$).

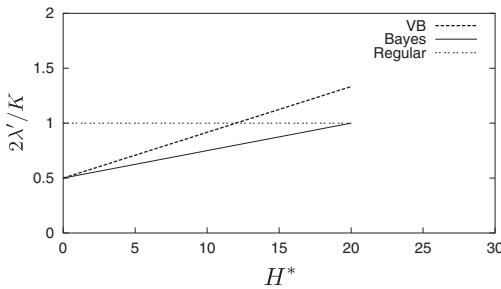


Figure 14.8 Free energy coefficients ($\max(L, M) = 50$, $\min(L, M) = 30$, $H = 20$, and $H^* = 1, \dots, 20$).

coefficients of the RRR model with the same setting as Figures 14.1 through 14.4, respectively. As for the generalization and the training coefficients, the vertical axis indicates the free energy coefficient normalized by the half of the *essential* parameter dimensionality D , given by Eq. (14.4). The curves correspond to the VB free energy coefficient (Theorem 14.18), the Bayes free energy coefficient (Theorem 14.23), and the Bayes free energy coefficient $2\lambda'_{\text{Regular}}^{\text{Bayes}} = D$ of the regular models with the same parameter dimensionality. We find that the VB free energy almost coincides with the Bayes free energy in Figures 14.5 and 14.6, while the VB free energy is much larger than the Bayes free energy in Figures 14.7 and 14.8.

Since the gap between the VB free energy and the Bayes free energy indicates how well the VB posterior approximates the Bayes posterior in terms of the KL divergence (see Section 13.6), our observation is not exactly what we would expect. For example, we see in Figure 14.1 that the generalization performance of VB learning is significantly different from Bayesian learning (when $H \ll \min(L, M)$ and when $H \sim \min(L, M)$), while the free energies in

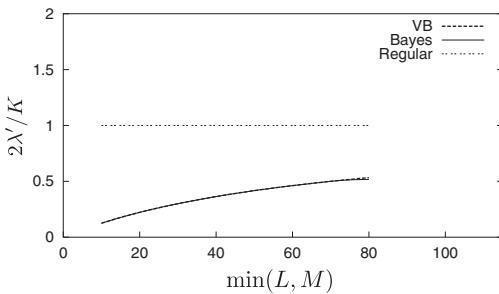


Figure 14.9 Free energy coefficients ($\max(L, M) = 80$, $\min(L, M) = 10, \dots, 80$, $H = 10$, and $H^* = 0$).

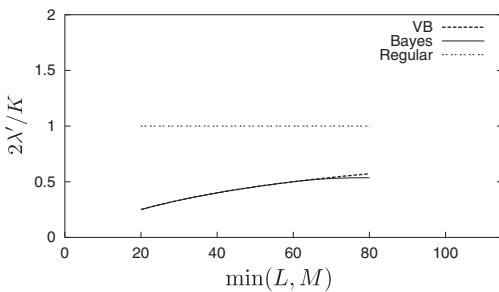


Figure 14.10 Free energy coefficients ($\max(L, M) = 80$, $\min(L, M) = 20, \dots, 80$, $H = 20$, and $H^* = 0$).

Figure 14.5 imply that the VB posterior well approximates the Bayes posterior. Also, by comparing Figures 14.3 and 14.7, we observe that, when $H \ll \min(L, M)$, VB learning provides much worse generalization performance than Bayesian learning, while the VB free energy well approximates the Bayes free energy; and that, when $H \sim \min(L, M)$, VB learning provides much better generalization performance, while the VB free energy is significantly larger than the Bayes free energy. Further investigation is required to understand the relation between the generalization performance and the gap between the VB and the Bayes free energies.

Figures 14.9 through 14.11 show similar cases to Figure 14.6 but for different ranks $H = 10, 20, 40$, respectively. From Figures 14.5 through 14.11, we conclude that, in general, the VB free energy behaves similarly to the Bayes free energy when L and M are significantly different from each other or $H \ll \min(L, M)$. In Figure 14.8, the VB free energy behaves strangely and poorly approximates the Bayes free energy when H^* is large. This is because

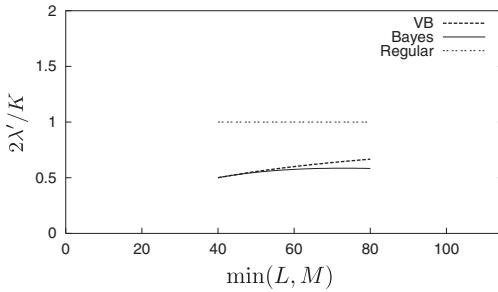


Figure 14.11 Free energy coefficients ($\max(L, M) = 80$, $\min(L, M) = 40, \dots, 80$, $H = 40$, and $H^* = 0$).

of the trivial redundancy of the RRR model, of which VB learning with the independence constraint cannot make use to reduce the free energy (see the remark in the last paragraph of Section 14.2.5).

14.2.7 Analysis with Subtle True Singular Values

Here we conduct an additional analysis to explain the seemingly contradictory observation in Figure 14.4—in the RRR model with $\max(L, M) = 50$, $\min(L, M) = 30$, $H = 20$, VB learning always gives smaller generalization error than Bayesian learning, regardless of the true rank H^* . We show that this does not mean the domination by VB learning over Bayesian learning, which was proven to be never dominated by any other method (see Appendix D).

Distinct and Subtle Signal Assumptions

The contradictory observation was due to the assumption (14.42) on the true singular values:

$$\gamma_h^* = \begin{cases} \Theta(1) & \text{for } h = 1, \dots, H^*, \\ 0 & \text{for } h = H^* + 1, \dots, \min(L, M), \end{cases} \quad (14.86)$$

which we call the *distinct signal assumption*. This assumption seems to cover any true linear mapping $\mathbf{B}^* \mathbf{A}^{*\top} = \sum_{h=1}^H \gamma_h^* \omega_{b_h}^* \omega_{a_h}^*$ by classifying all singular components such that $\gamma_h^* > 0$ to the necessary components $h = 1, \dots, H^*$, and the other components such that $\gamma_h^* = 0$ to the redundant components $h = H^* + 1, \dots, \min(L, M)$. However, in the asymptotic limit, the assumption (14.86) *implicitly* prohibits the existence of true singular values in the same order as the noise contribution, i.e., $\gamma_h^* = \Theta_p(N^{-1/2})$. In other words, the distinct signal assumption (14.86) considers all true singular values to be either *infinitely*

larger than the noise or exactly equal to zero. As a result, asymptotic analysis under the distinct signal assumption reflects only the *overfitting* tendency of a learning machine, and ignores the *underfitting* tendency, which happens when the signal is not clearly separable from the noise. Since overfitting and underfitting are in the trade-off relation, it is important to investigate both tendencies when generalization performance is analyzed.

To relax the restriction discussed previously, we replace the assumption (14.42) with

$$\gamma_h^* = \begin{cases} \Theta(1) & \text{for } h = 1, \dots, H^*, \\ O(N^{-1/2}) & \text{for } h = H^* + 1, \dots, \min(L, M), \end{cases} \quad (14.87)$$

which we call the *subtle signal assumption*, in the following analysis (Watanabe and Amari, 2003; Nakajima and Watanabe, 2007). Note that, with the assumption (14.87), we do not intend to analyze the case where the true singular values depend on N . Rather, we assume realistic situations where the number of necessary components H^* depends on N . Let us keep in mind the following two points, which are usually true when we analyze real-world data:

- The number N of samples is always finite.

Asymptotic theory is not to investigate what happens when $N \rightarrow \infty$, but to approximate the situation where N is finite but large.

- It rarely happens that real-world data can be *exactly* expressed by a low-rank model.

Statistical models are supposed to be simpler than the real-world data generation process, but expected to approximate it with certain accuracy, and the accuracy depends on the noise level and the number of samples.

Then we expect that, for most real-world data, it holds that $\gamma_h^* > 0$ for all $h = 1, \dots, \min(L, M)$, but, given finite N , some of the true singular values are comparable to the noise contribution $\gamma_h^* = \Theta(N^{-1/2})$, and some others are negligible $\gamma_h^* = o(N^{-1/2})$. The subtle signal assumption (14.87) covers such realistic situations.

Generalization Error under Subtle Signal Assumption

Replacing the distinct signal assumption (14.86) with the subtle signal assumption (14.87) does not affect the discussion up to Theorem 14.10, i.e., Theorems 14.1, 14.5, and 14.10, Lemmas 14.6 through 14.8, and their corollaries still hold. Instead of Theorem 14.11, we have the following theorem:

Theorem 14.25 *Under the subtle signal assumption (14.87), the average generalization error of the RRR model for $H \geq H^*$ is asymptotically expanded as*

$$\overline{\text{GE}}(N) = \lambda^{\text{VB}} N^{-1} + O(N^{-3/2}),$$

where the generalization coefficient is given by

$$\begin{aligned} 2\lambda^{\text{VB}} &= (H^*(L+M) - H^{*2}) + \frac{N}{\sigma'^2} \sum_{h=H^*+1}^{\min(L,M)} \gamma_h'^* \\ &+ \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^* > \max(L, M)) \right. \\ &\cdot \left. \left\{ \left(1 - \frac{\max(L,M)}{\gamma_h'^2}\right)^2 \gamma_h'^* - 2 \left(1 - \frac{\max(L,M)}{\gamma_h'^2}\right) \gamma_h'' \omega_{b_h}^{\prime\prime\top} V'''^* \omega_{a_h}'' \right\} \right\rangle_{q(V'')} . \end{aligned} \quad (14.88)$$

Here,

$$V'' = \sum_{h=1}^{\min(L,M)-H^*} \gamma_h'' \omega_{b_h}'' \omega_{a_h}^{\prime\prime\top} \quad (14.89)$$

is the SVD of a random matrix $V'' \in \mathbb{R}^{(\min(L,M)-H^*) \times (\max(L,M)-H^*)}$ subject to

$$q(V'') = \text{MGauss}_{\min(L,M)-H^*, \max(L,M)-H^*}(V''; V''^*, I_{\min(L,M)-H^*} \otimes I_{\max(L,M)-H^*}), \quad (14.90)$$

and $V''^* \in \mathbb{R}^{(\min(L,M)-H^*) \times (\max(L,M)-H^*)}$ is a (nonsquare) diagonal matrix with the diagonal entries given by $V_{h,h}'' = \frac{\sqrt{N}}{\sigma'} \gamma_{H^*+h}^*$ for $h = 1, \dots, \min(L, M) - H^*$.

Proof From Eq. (14.58), we have

$$\begin{aligned} \|\widehat{BA}^\top - B^* A^{*\top}\|_{\text{Fro}}^2 &= \left\| \sum_{h=1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top - B^* A^{*\top} \right\|_{\text{Fro}}^2 \\ &= \left\| \sum_{h=1}^{H^*} \gamma_h \omega_{b_h} \omega_{a_h}^\top - B^* A^{*\top} + \sum_{h=H^*+1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top \right\|_{\text{Fro}}^2 + O_p(N^{-3/2}) \\ &= \left\| V - B^* A^* + \sum_{h=H^*+1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top - \sum_{h=H^*+1}^{\min(L,M)} \gamma_h \omega_{b_h} \omega_{a_h}^\top \right\|_{\text{Fro}}^2 \\ &\quad + O_p(N^{-3/2}), \end{aligned}$$

and therefore,

$$\begin{aligned} &\left\langle \|\widehat{BA}^\top - B^* A^{*\top}\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\ &= \left\langle \left\| V - B^* A^* + \sum_{h=H^*+1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top - \sum_{h=H^*+1}^{\min(L,M)} \gamma_h \omega_{b_h} \omega_{a_h}^\top \right\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\ &\quad + O(N^{-3/2}) \\ &= \left\langle \|V - B^* A^*\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\ &\quad + 2 \left\langle (V - B^* A^*)^\top \left(\sum_{h=H^*+1}^H \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top - \sum_{h=H^*+1}^{\min(L,M)} \gamma_h \omega_{b_h} \omega_{a_h}^\top \right) \right\rangle_{q(\mathcal{D})} \end{aligned}$$

$$\begin{aligned}
& + \left\langle \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}})^2 - 2 \sum_{h=H^*+1}^H \gamma_h \widehat{\gamma}_h^{\text{VB}} + \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^2 \right\rangle_{q(\mathcal{D})} \\
& + O(N^{-3/2}) \\
& = \left\langle \|V - B^* A^*\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\
& + 2 \left\langle \sum_{h=H^*+1}^H (\gamma_h \omega_{b_h} \omega_{a_h}^\top - \gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top})^\top \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top \right. \\
& \quad \left. - \sum_{h=H^*+1}^{\min(L,M)} (\gamma_h \omega_{b_h} \omega_{a_h}^\top - \gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top})^\top \gamma_h \omega_{b_h} \omega_{a_h}^\top \right\rangle_{q(\mathcal{D})} \\
& + \left\langle \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}})^2 - 2 \sum_{h=H^*+1}^H \gamma_h \widehat{\gamma}_h^{\text{VB}} + \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^2 \right\rangle_{q(\mathcal{D})} \\
& + O(N^{-3/2}) \\
& = \left\langle \|V - B^* A^*\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} - 2 \left\langle \sum_{h=H^*+1}^H (\gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top})^\top \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top \right\rangle_{q(\mathcal{D})} \\
& + 2 \left\langle \sum_{h=H^*+1}^{\min(L,M)} (\gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top})^\top \gamma_h \omega_{b_h} \omega_{a_h}^\top \right\rangle_{q(\mathcal{D})} \\
& + \left\langle \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}})^2 \right\rangle_{q(\mathcal{D})} - \left\langle \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^2 \right\rangle_{q(\mathcal{D})} + O(N^{-3/2}) \\
& = \left\langle \|V - B^* A^*\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} - \left\langle \sum_{h=H^*+1}^{\min(L,M)} \left\| \gamma_h \omega_{b_h} \omega_{a_h}^\top - \gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top} \right\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\
& - 2 \left\langle \sum_{h=H^*+1}^H (\gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top})^\top \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top \right\rangle_{q(\mathcal{D})} \\
& + \left\langle \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}})^2 \right\rangle_{q(\mathcal{D})} + \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^{*2} + O(N^{-3/2}) \\
& = \frac{\sigma'^2}{N} \left(LM - (L - H^*)(M - H^*) + \frac{N}{\sigma'^2} \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^{*2} \right) \\
& + \left\langle \sum_{h=H^*+1}^H (\widehat{\gamma}_h^{\text{VB}})^2 \right\rangle_{q(\mathcal{D})} - 2 \left\langle \sum_{h=H^*+1}^H (\gamma_h^* \omega_{b_h}^* \omega_{a_h}^{*\top})^\top \widehat{\gamma}_h^{\text{VB}} \omega_{b_h} \omega_{a_h}^\top \right\rangle_{q(\mathcal{D})} \\
& + O(N^{-3/2}). \tag{14.91}
\end{aligned}$$

In the orthogonal space to the *distinctly* necessary components $\{\gamma_h, \omega_{a_h}, \omega_{b_h}\}_{h=1}^{H^*}$, the distribution of $\{\gamma_h, \omega_{a_h}, \omega_{b_h}\}_{h=H^*+1}^{\min(L,M)}$ coincides with the distribution of $\{\frac{\sigma'^2}{\sqrt{N}} \gamma_h'', \omega_{a_h}'', \omega_{b_h}''\}_{h=1}^{\min(L,M)-H^*}$, defined in Eq. (14.89), with V'''^* as the true matrix for subtle or the redundant components, $h = H^* + 1, \dots, \min(L, M)$. By using Eq. (14.58), we thus have

$$\begin{aligned}
& \left\langle \left\| \widehat{BA}^\top - B^* A^{*\top} \right\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\
& = \frac{\sigma'^2}{N} \left((H^*(L + M) - H^{*2}) + \frac{N}{\sigma'^2} \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^{*2} \right. \\
& \quad \left. + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > \max(L, M)) \right\rangle \right)
\end{aligned}$$

$$\begin{aligned} & \cdot \left\{ \left(1 - \frac{\max(L, M)}{\gamma_h'^2} \right)^2 \gamma_h'^2 - 2 \left(1 - \frac{\max(L, M)}{\gamma_h'^2} \right) \gamma_h'' \omega_{b_h}^{\prime\prime\top} V'''^* \omega_{a_h}'' \right\}_{q(V''')} \\ & + O(N^{-3/2}), \end{aligned}$$

which completes the proof. \square

Training Error under Subtle Signal Assumption

The training error can be analyzed more easily.

Theorem 14.26 *Under the subtle signal assumption (14.87), the average training error of the RRR model for $H \geq H^*$ is asymptotically expanded as*

$$\overline{\text{TE}}(N) = \nu^{\text{VB}} N^{-1} + O(N^{-3/2}),$$

where the training coefficient is given by

$$\begin{aligned} 2\nu^{\text{VB}} = & -(H^*(L+M) - H^{*2}) + \frac{N}{\sigma'^2} \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^{*2} \\ & + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > \max(L, M)) \cdot \left(1 - \frac{\max(L, M)}{\gamma_h'^2} \right) \left(1 + \frac{\max(L, M)}{\gamma_h'^2} \right) \gamma_h'^2 \right\rangle_{q(V''')} . \end{aligned} \quad (14.92)$$

Here V'' and $\{\gamma_h''\}$ are defined in Theorem 14.25.

Proof Theorem 14.15 and Eq. (14.77) still hold under the assumption (14.87). Therefore,

$$\begin{aligned} & \|V - \widehat{BA}^\top\|_{\text{Fro}}^2 \\ &= - \sum_{h=H^*+1}^H \theta\left(\gamma_h^2 > \frac{\max(L, M)\sigma'^2}{N}\right) \cdot \left(\gamma_h - \frac{\max(L, M)\sigma'^2}{N\gamma_h}\right) \left(\gamma_h + \frac{\max(L, M)\sigma'^2}{N\gamma_h}\right) \\ &+ \sum_{h=H^*+1}^{\min(L,M)} (\gamma_h - \gamma_h^*)^2 + \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^{*2} + O_p(N^{-3/2}), \end{aligned}$$

and

$$\begin{aligned} & \left\langle \|V - \widehat{BA}^\top\|_{\text{Fro}}^2 - \|V - B^* A^{*\top}\|_{\text{Fro}}^2 \right\rangle_{q(\mathcal{D})} \\ &= -\frac{\sigma'^2}{N} \left\{ (H^*(L+M) - H^{*2}) - \frac{N}{\sigma'^2} \sum_{h=H^*+1}^{\min(L,M)} \gamma_h^{*2} \right. \\ &+ \sum_{h=H^*+1}^H \theta\left(\gamma_h^2 > \frac{\max(L, M)\sigma'^2}{N}\right) \cdot \left(\gamma_h - \frac{\max(L, M)\sigma'^2}{N\gamma_h}\right) \left(\gamma_h + \frac{\max(L, M)\sigma'^2}{N\gamma_h}\right) \Big\} \\ &+ O_p(N^{-3/2}). \end{aligned}$$

Substituting the preceding equation into Eq. (14.75) and using the random matrix V'' and its singular values $\{\gamma_h''\}$, defined in Theorem 14.25, we obtain Eq. (14.92). \square

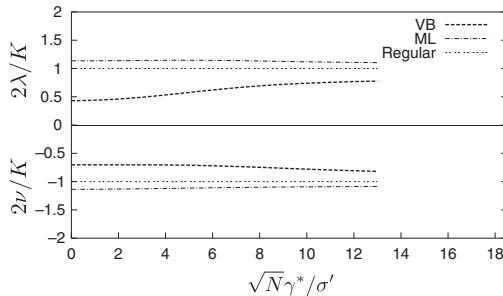


Figure 14.12 The generalization coefficients and the training coefficients under the subtle signal assumption (14.87) in the RRR model with $\max(L, M) = 50$, $\min(L, M) = 30$, $H = 20$, and $H^* = 5$.

Comparison with Other Learning Algorithms

Figure 14.12 shows the generalization coefficients and the training coefficients, computed by using Theorems 14.25 and 14.26, respectively, as functions of a rescaled subtle true singular value $\sqrt{N}\gamma_h^*/\sigma'$. The considered RRR model is with $\max(L, M) = 50$, $\min(L, M) = 30$, and $H = 20$, and the true linear mapping is assumed to consist of $H^* = 5$ *distinctly* necessary components ($\gamma_h^* = \Theta(1)$ for $h = 1, \dots, 5$), 10 *subtle* components ($\gamma_h^* = \Theta(N^{-1/2})$ for $h = 6, \dots, 15$), and the other five null components ($\gamma_h^* = 0$ for $h = 16, \dots, 20$). The subtle singular values are assumed to be identical, $\gamma_h^* = \gamma^*$ for $h = 6, \dots, 15$, and the horizontal axis indicates $\sqrt{N}\gamma^*/\sigma'$. The generalization coefficients and the training coefficients of ML learning can be derived in the same way as Theorems 14.25 and 14.26 with the VB estimator $\widehat{\gamma}_h^{\text{VB}}$ replaced with the ML estimator $\widehat{\gamma}_h^{\text{ML}} = \gamma_h$. Unfortunately, the generalization error nor the training error of Bayesian learning under the subtle signal assumption for the general RRR model has not been clarified.

Only in the case where $L = H = 1$, the Bayes generalization error under the subtle signal assumption has been analyzed.

Theorem 14.27 (Watanabe and Amari, 2003) *The Bayes generalization error of the RRR model with $M \geq 2, L = H = 1$ under the assumption that the true mapping is $b^* \mathbf{a}^* = O(N^{-1/2})$ is asymptotically expanded as*

$$\overline{\text{GE}}^{\text{Bayes}}(N) = \lambda^{\text{Bayes}} N^{-1} + o(N^{-1}),$$

where the generalization coefficient is given by

$$2\lambda^{\text{Bayes}} = 1 + \left\langle \left(\left\| \frac{\sqrt{N}}{\sigma'} b^* \mathbf{a}^* \right\|^2 + \frac{\sqrt{N}}{\sigma'} b^* \mathbf{a}^{*\top} \mathbf{v} \right) \frac{\Phi_M(\mathbf{v})}{\Phi_{M-2}(\mathbf{v})} \right\rangle_{q(\mathbf{v})}. \quad (14.93)$$

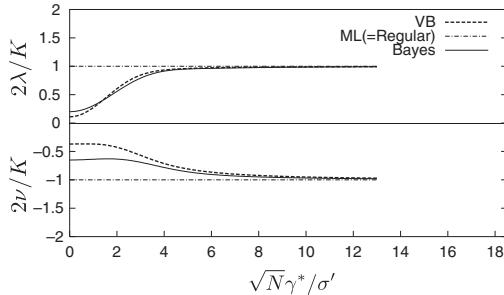


Figure 14.13 The generalization coefficients and the training coefficients under the subtle signal assumption (14.87) in the RRR model with $M = 5$, $L = H = 1$, and $H^* = 0$.

Here,

$$\Phi_M(\mathbf{v}) = \int_0^{\pi/2} \sin^M \theta \exp\left(-\frac{1}{2} \left\| \frac{\sqrt{N}}{\sigma'} b^* \mathbf{a}^* + \mathbf{v} \right\|^2 \sin^2 \theta\right) d\theta,$$

and $\mathbf{v} \in \mathbb{R}^M$ is a random vector subject to $q(\mathbf{v}) = \text{Gauss}_M(\mathbf{v}; \mathbf{0}, \mathbf{I}_M)$.

Figure 14.13 compares the generalization coefficients when $M = 5$, $L = H = 1$, and $H^* = 0$, where the horizontal axis corresponds to a rescaled subtle true singular value $\sqrt{N}\gamma^*/\sigma' = \sqrt{N}\|b^*\mathbf{a}^*\|/\sigma'$.² We see that the generalization error of VB learning is smaller than that of Bayesian learning when $\sqrt{N}\gamma^*/\sigma' = 0$, and identical when $\sqrt{N}\gamma^*/\sigma' \rightarrow \infty$. This means that, under the distinct signal assumption (14.86), which considers only the case where $\gamma^* = 0$ (i.e., $H^* = 0$) or $\gamma^* = \Theta(1)$ (i.e., $H^* = 1$), VB learning always performs better than Bayesian learning. However, we can see in Figure 14.13 that, when $\sqrt{N}\gamma^*/\sigma' \approx 3$, Bayesian learning outperforms VB learning. Figure 14.13 simply implies that VB learning is more strongly regularized than Bayesian learning, or in other words, VB learning tends to underfit subtle signals such that $\gamma^* = \Theta(N^{-1/2})$, while Bayesian learning tends to overfit noise.

Knowing the proved optimality of Bayesian learning (Appendix D), we would expect that the same happens in Figure 14.12, where the limits $\sqrt{N}\gamma^*/\sigma' = 0$ and $\sqrt{N}\gamma^*/\sigma' \rightarrow \infty$ correspond to the cases with $H^* = 5$ and $H^* = 15$, respectively, in Figure 14.4 under the distinct signal assumption

² When $L = H = 1$, the parameter transformation $b\mathbf{a} \rightarrow \mathbf{w}$ makes the RRR model identifiable, and therefore the ML generalization coefficient is identical to that of the regular models. This is the reason why only the *integration effect* or *model induced-regularization* was observed in the one-dimensional matrix factorization model in Section 7.2 and Section 7.3.3. The *basis selection effect* appears only when a singular model cannot be equivalently transformed to a regular model (see the discussion in Section 14.3).

(14.86). Namely, if we could depict the Bayes generalization coefficient in Figure 14.12, there should be some interval where Bayesian learning outperforms VB learning.

14.3 Insights into VB Learning

In this chapter, we analyzed the generalization error, the training error, and the free energy of VB learning in the RRR model, and derived their asymptotic forms. We also introduced theoretical results providing those properties for ML learning and Bayesian learning. As mentioned in Section 13.5, the RRR model is the only singular model of which those three properties have been theoretically clarified for ML learning, Bayesian learning, and VB learning. Accordingly, we here summarize our observations, and discuss effects of singularities in VB learning and other learning algorithms.

- (i) In the RRR model, the *basis selection effect*, explained in Section 13.5.1, appears as a selection bias of largest singular values of a zero-mean random matrix.

Theorem 14.19 gives an asymptotic expansion of the ML generalization error. The second term in the generalization coefficient (14.82) is the expectation of the square of the $(H - H^*)$ largest singular values of a random matrix $\frac{\sqrt{N}}{\sigma'} \mathbf{V}'$, where \mathbf{V}' is subject to the zero-mean Gaussian (14.61). This corresponds to the effect of basis selection: ML learning chooses the singular components that best fit the observation noise. With the full-rank model, i.e., $H = \min(L, M)$, the second term in the generalization coefficient (14.82) is equal to

$$\begin{aligned} \left\langle \sum_{h=1}^{\min(L,M)-H^*} \gamma_h'^2 \right\rangle_{q(\mathbf{W})} &= (\min(L, M) - H^*)(\max(L, M) - H^*) \\ &= (L - H^*)(M - H^*), \end{aligned}$$

and therefore the generalization coefficient becomes

$$\begin{aligned} 2\lambda^{\text{ML}} &= (H^*(L + M) - H^{*2}) + (L - H^*)(M - H^*) = LM \\ &= D, \end{aligned}$$

which is the same as the generalization coefficient of the regular models. Indeed, the full-rank RRR model is equivalently transformed to a regular model by $\mathbf{B}\mathbf{A}^\top \rightarrow \mathbf{U}$, where the domain for \mathbf{U} is the whole $\mathbb{R}^{L \times M}$ space (no low-rank restriction is imposed to \mathbf{U}). In this case, no *basis selection* occurs because all possible bases are supposed to be used.

- (ii) In the RRR model, the integration effect, explained in Section 13.5.1, appears as the James–Stein (JS) type shrinkage.

This was shown in Theorem 14.1 in the asymptotic limit: the VB estimator converges to the positive-part JS estimator operated on each singular component separately. By comparing Theorems 14.11 and 14.19, we see that ML learning and VB learning differ from each other by the factor $\theta(\gamma_h^2 > \max(L, M)) \left(1 - \frac{\max(L, M)}{\gamma_h^2}\right)^2$, which comes from the positive-part JS shrinkage. Unlike the basis selection effect, the integration effect appears even if the model can be equivalently transformed to a regular model—the full-rank RRR model (with $H = \min(L, M)$) is still affected by the singularities. The relation between VB learning and the JS shrinkage estimator was also observed in nonasymptotic analysis in Chapter 7, where model-induced regularization (MIR) was illustrated as a consequence of the integration effect, by focusing on the one-dimensional matrix factorization model. Note that basis selection effect does not appear in the one-dimensional matrix factorization model (where $L = M = H$), because it can be equivalently transformed to a regular model.

- (iii) VB learning shows similarity both to ML learning and Bayesian learning.

Figures 14.1 through 14.4 generally show that VB learning is regularized as much as Bayesian learning, while its dependence on the model size (H, L, M , etc.) is more like ML learning. Unlike Bayesian learning, the integration effect does not always dominate the basis selection effect in VB learning—a good property of Bayesian learning, $2\lambda^{\text{Bayes}} \leq D$, does not necessarily hold in VB learning, e.g., we observe that $2\lambda^{\text{VB}} > D$ at $\min(L, M) = 80$ in Figure 14.2, and $H = 1$ in Figure 14.3.

- (iv) In VB learning, the relation between the generalization error and the free energy is not as simple as in Bayesian learning.

In Bayesian learning, the generalization coefficient and the free energy coefficient coincide with each other, i.e., $\lambda^{\text{Bayes}} = \lambda'^{\text{Bayes}}$. This property does not hold in VB learning even approximately, as seen by comparing Figures 14.1 through 14.4 and Figures 14.5 through 14.8. In many cases, the VB free energy well approximates the Bayes free energy, while the VB generalization error significantly differs from the Bayes generalization error. Research on the relation between the free energy and the generalization error in VB learning is ongoing (see Section 17.4).

- (v) MIR in VB learning can be stronger than that in Bayesian learning.

By definition, the VB free energy is never less than the Bayes free energy, and therefore it holds that $\lambda^{\text{VB}} \geq \lambda^{\text{Bayes}}$. On the other hand, such a relation does not hold for the generalization error, i.e., λ^{VB} can be larger or less than λ^{Bayes} . However, even if λ^{VB} is less than or equal to λ^{Bayes} for any true rank H^* in some RRR model, it does not mean the domination of VB learning over Bayesian learning. Since the optimality of Bayesian learning was proved (see Appendix D), $\lambda^{\text{VB}} < \lambda^{\text{Bayes}}$ simply means that VB learning is more strongly regularized than Bayesian learning, or in other words, VB learning tends to underfit small signals while Bayesian learning tends to overfit noise. Extending the analysis under the subtle signal assumption (14.87) to the general RRR model would clarify this point.

- (vi) The generalization error depends on the dimensionality in an interesting way.

The shrinkage factor is governed by $\max(L, M)$ and independent of $\min(L, M)$ in the asymptotic limit (see Theorem 14.1). This is because the shrinkage is caused by the VB posterior extending into the parameter space with larger dimensional space (M -dimensional input space or L -dimensional output space) for the redundant components, as seen in Corollary 14.3. This choice was made by maximizing the entropy of the VB posterior distribution when the free energy is minimized.

Consequently, when $L \neq M$, the shape of the VB posterior in the asymptotic limit is similar to the partially Bayesian learning, where the posterior of \mathbf{A} or \mathbf{B} is approximated by the Dirac delta function (see Chapter 12). On the other hand, increase of the smaller dimensionality $\min(L, M)$ broadens the variety of basis selection: as mentioned in (i), the basis selection effect in the RRR model occurs by the redundant components selecting the $(H - H^*)$ largest singular components, and $(\min(L, M) - H^*)$ corresponds to the dimensionality that the basis functions can span. This phenomenon can be seen in Figure 8.3—the Marčenko–Pastur distribution is diverse when $\alpha = (\min(L, M) - H^*) / (\max(L, M) - H^*)$ is large. We can conclude that a large $\max(L, M)$ enhances the integration effect, leading to strong regularization, while a large $\min(L, M)$ enhances the basis selection effect, leading to overfitting. As a result, VB learning tends to be strongly regularized when $L \ll M$ or $L \gg M$, and tends to overfit when $L \approx M$.

15

Asymptotic VB Theory of Mixture Models

In this chapter, we discuss the asymptotic behavior of the VB free energy of mixture models, for which VB learning algorithms were introduced in Sections 4.1.1 and 4.1.2. We first prepare basic lemmas commonly used in this and the following chapters.

15.1 Basic Lemmas

Consider the *latent variable model* expressed as

$$p(\mathcal{D}|\mathbf{w}) = \sum_{\mathcal{H}} p(\mathcal{D}, \mathcal{H}|\mathbf{w}).$$

In this chapter, we analyze the VB free energy, which is the minimum of the free energy under the constraint,

$$r(\mathbf{w}, \mathcal{H}) = r_w(\mathbf{w})r_{\mathcal{H}}(\mathcal{H}), \quad (15.1)$$

i.e.,

$$F^{\text{VB}}(\mathcal{D}) = \min_{r_w(\mathbf{w}), r_{\mathcal{H}}(\mathcal{H})} F(r), \quad (15.2)$$

where

$$\begin{aligned} F(r) &= \left\langle \log \frac{r_w(\mathbf{w})r_{\mathcal{H}}(\mathcal{H})}{p(\mathbf{w}, \mathcal{H}, \mathcal{D})} \right\rangle_{r_w(\mathbf{w})r_{\mathcal{H}}(\mathcal{H})} \\ &= F^{\text{Bayes}}(\mathcal{D}) + \text{KL}(r_w(\mathbf{w})r_{\mathcal{H}}(\mathcal{H}) \| p(\mathbf{w}, \mathcal{H}|\mathcal{D})) . \end{aligned} \quad (15.3)$$

Here,

$$F^{\text{Bayes}}(\mathcal{D}) = -\log p(\mathcal{D}) = -\log \sum_{\mathcal{H}} p(\mathcal{D}, \mathcal{H}) = -\log \sum_{\mathcal{H}} \int p(\mathcal{D}, \mathcal{H}, \mathbf{w}) d\mathbf{w}$$

is the Bayes free energy. Recall that the stationary condition of the free energy yields

$$r_w(\mathbf{w}) = \frac{1}{C_w} p(\mathbf{w}) \exp \langle \log p(\mathcal{D}, \mathcal{H}|\mathbf{w}) \rangle_{r_{\mathcal{H}}(\mathcal{H})}, \quad (15.4)$$

$$r_{\mathcal{H}}(\mathcal{H}) = \frac{1}{C_{\mathcal{H}}} \exp \langle \log p(\mathcal{D}, \mathcal{H}|\mathbf{w}) \rangle_{r_w(\mathbf{w})}. \quad (15.5)$$

For the minimizer $\widehat{r}_w(\mathbf{w})$ of $F(r)$, let

$$\widehat{\mathbf{w}} = \langle \mathbf{w} \rangle_{\widehat{r}_w(\mathbf{w})} \quad (15.6)$$

be the VB estimator.

The following lemma shows that the free energy is decomposed into the sum of two terms.

Lemma 15.1 *It holds that*

$$F^{\text{VB}}(\mathcal{D}) = \min_{r_w(\mathbf{w})} \{R + Q\}, \quad (15.7)$$

where

$$\begin{aligned} R &= \text{KL}(r_w(\mathbf{w}) \| p(\mathbf{w})), \\ Q &= -\log C_{\mathcal{H}}, \end{aligned}$$

for $C_{\mathcal{H}} = \sum_{\mathcal{H}} \exp \langle \log p(\mathcal{D}, \mathcal{H}|\mathbf{w}) \rangle_{r_w(\mathbf{w})}$.

Proof From the restriction of the VB approximation in Eq. (15.1), $F(r)$ can be divided into two terms,

$$F(r) = \left\langle \log \frac{r_w(\mathbf{w})}{p(\mathbf{w})} \right\rangle_{r_w(\mathbf{w})} + \left\langle \log \frac{r_{\mathcal{H}}(\mathcal{H})}{p(\mathcal{D}, \mathcal{H}|\mathbf{w})} \right\rangle_{r_w(\mathbf{w}) r_{\mathcal{H}}(\mathcal{H})}.$$

Since the optimal VB posteriors satisfy Eqs. (15.4) and (15.5), if the VB posterior $r_{\mathcal{H}}(\mathcal{H})$ is optimized, then

$$\left\langle \log \frac{r_{\mathcal{H}}(\mathcal{H})}{p(\mathcal{D}, \mathcal{H}|\mathbf{w})} \right\rangle_{r_w(\mathbf{w}) r_{\mathcal{H}}(\mathcal{H})} = -\log C_{\mathcal{H}}$$

holds. Thus, we obtain Eq. (15.7). \square

The free energies of mixture models and other latent variable models involve the di-gamma function $\Psi(x)$ and the log-gamma function $\log \Gamma(x)$ (see, e.g., Eq. (4.22)). To analyze the free energy, we will use the inequalities on these functions in the following lemma:

Lemma 15.2 (Alzer, 1997) For $x > 0$,

$$\frac{1}{2x} < \log x - \Psi(x) < \frac{1}{x}, \quad (15.8)$$

and

$$0 \leq \log \Gamma(x) - \left\{ \left(x - \frac{1}{2} \right) \log x - x + \frac{1}{2} \log 2\pi \right\} \leq \frac{1}{12x}. \quad (15.9)$$

The inequalities (15.8) ensure that substituting $\log x$ for $\Psi(x)$ only contributes at most additive constant terms to the VB free energy. The substitution for $\log \Gamma(x)$ is given by Eq. (15.9) as well.

For the i.i.d. latent variable models defined as

$$p(\mathbf{x}|\mathbf{w}) = \sum_z p(\mathbf{x}, z|\mathbf{w}), \quad (15.10)$$

the likelihood for the observed data $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^N$ and the complete data $\{\mathcal{D}, \mathcal{H}\} = \{\mathbf{x}^{(n)}, \mathbf{z}^{(n)}\}_{n=1}^N$ is given by

$$\begin{aligned} p(\mathcal{D}|\mathbf{w}) &= \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}), \\ p(\mathcal{D}, \mathcal{H}|\mathbf{w}) &= \prod_{n=1}^N p(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}|\mathbf{w}), \end{aligned}$$

respectively. In the asymptotic analysis of the free energy for such a model, when the free energy is minimized, the second term in Eq. (15.7), $Q = -\log C_{\mathcal{H}}$, is proved to be close to N times the empirical entropy (13.20),

$$S_N(\mathcal{D}) = -\frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}^{(n)}|\mathbf{w}^*), \quad (15.11)$$

where \mathbf{w}^* is the true parameter generating the data. Thus, the first term in Eq. (15.7) shows the asymptotic behavior of the VB free energy, which is analyzed with the inequalities in Lemma 15.2.

Let

$$\tilde{Q} = Q - NS_N(\mathcal{D}). \quad (15.12)$$

It follows from Jensen's inequality that

$$\begin{aligned} \tilde{Q} &= \log p(\mathcal{D}|\mathbf{w}^*) - \log \sum_{\mathcal{H}} \exp \langle \log p(\mathcal{D}, \mathcal{H}|\mathbf{w}) \rangle_{r_w(\mathbf{w})} \\ &\geq \log \frac{p(\mathcal{D}|\mathbf{w}^*)}{\langle p(\mathcal{D}|\mathbf{w}) \rangle_{r_w(\mathbf{w})}} \\ &\geq N E_N(\tilde{\mathbf{w}}^{\text{ML}}), \end{aligned} \quad (15.13)$$

where $\widehat{\mathbf{w}}^{\text{ML}}$ is the *maximum likelihood (ML) estimator*, and

$$\begin{aligned} E_N(\mathbf{w}) &= L_N(\mathbf{w}^*) - L_N(\mathbf{w}) \\ &= \frac{1}{N} \sum_{n=1}^N \log \frac{p(\mathbf{x}^{(n)}|\mathbf{w}^*)}{p(\mathbf{x}^{(n)}|\mathbf{w})} \end{aligned} \quad (15.14)$$

is the empirical KL divergence. Note here that L_N is defined in Eq. (13.37), and $E_N(\mathbf{w})$ corresponds to the training error of the plug-in predictive distribution (defined in Eq. (13.75) for regular models) with an estimator \mathbf{w} .

If the domain of data \mathcal{X} is discrete and with finite cardinality, $\#(\mathcal{X}) = M$, Q in Eq. (15.7) can be analyzed in detail. In such a case, we can assume without loss of generality that $\mathbf{x} \in \{\mathbf{e}_1, \dots, \mathbf{e}_M\}$, where \mathbf{e}_m is the one-of- M representation, i.e., only the m th entry is one and the other entries are zeros. Let \widehat{N}_m be the number of output m in the sequence \mathcal{D} , i.e., $\widehat{N}_m = \sum_{n=1}^N x_m^{(n)}$, and define the strongly ε -typical set $T_\varepsilon^N(\mathbf{p}^*)$ with respect to the probability mass function,

$$\mathbf{p}^* = (p_1^*, \dots, p_M^*)^\top = (p(x_1 = 1|\mathbf{w}^*), \dots, p(x_M = 1|\mathbf{w}^*))^\top \in \Delta^{M-1}$$

as follows:

$$T_\varepsilon^N(\mathbf{p}^*) = \left\{ \mathcal{D} \in \mathcal{X}^N; \left| \frac{\widehat{N}_m}{N} - p_m^* \right| \leq \frac{p_m^*}{\log M} \varepsilon, \quad m = 1, \dots, M \right\}. \quad (15.15)$$

It is known that the probability that the observed data sequence is not strongly ε -typical is upper-bounded as follows:

Lemma 15.3 (Han and Kobayashi, 2007) *It holds that*

$$\text{Prob}(\mathcal{D} \notin T_\varepsilon^N(\mathbf{p}^*)) \leq \frac{\kappa M}{N\varepsilon^2},$$

where

$$\kappa = (\log M)^2 \max_{m:p_m^* \neq 0} \frac{1 - p_m^*}{p_m^*}.$$

Let

$$\widehat{\mathbf{p}} = (\widehat{p}_1, \dots, \widehat{p}_M)^\top = \left(\langle p(x_1 = 1|\mathbf{w}) \rangle_{r_w(\mathbf{w})}, \dots, \langle p(x_M = 1|\mathbf{w}) \rangle_{r_w(\mathbf{w})} \right)^\top \in \Delta^{M-1}$$

be the probability mass function defined by the predictive distribution $\langle p(\mathbf{x}|\mathbf{w}) \rangle_{r_w(\mathbf{w})}$ with the VB posterior $r_w(\mathbf{w})$. For any fixed $\delta > 0$, define

$$R_\delta^* = \left\{ \widehat{\mathbf{p}} \in \Delta^{M-1}; \text{KL}(\mathbf{p}^* \parallel \widehat{\mathbf{p}}) \leq \delta \right\}, \quad (15.16)$$

where $\text{KL}(\mathbf{p}^* \parallel \widehat{\mathbf{p}}) = \sum_{m=1}^M p_m^* \log \frac{p_m^*}{\widehat{p}_m}$. Then the following lemma holds:

Lemma 15.4 Suppose that the domain \mathcal{X} is discrete and with finite cardinality, $\#(\mathcal{X}) = M$. For all $\varepsilon > 0$ and $\mathcal{D} \in T_\varepsilon^N(\mathbf{p}^*)$, there exists a constant $C > 0$ such that if $\widehat{\mathbf{p}} \notin R_{C\varepsilon^2}^*$,

$$\widetilde{Q} = \Omega_p(N). \quad (15.17)$$

Furthermore,

$$\min_{r_w(\mathbf{w})} \widetilde{Q} = O_p(1). \quad (15.18)$$

Proof From Eq. (15.13), we have

$$\begin{aligned} \widetilde{Q} &\geq \log \frac{p(\mathcal{D}|\mathbf{w}^*)}{\langle p(\mathcal{D}|\mathbf{w}) \rangle_{r_w(\mathbf{w})}} \\ &= N \sum_{m=1}^M \frac{\widehat{N}_m}{N} \log \frac{p_m^*}{\widehat{p}_m} \\ &= N \left\{ \text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \widehat{\mathbf{p}}) - \text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \mathbf{p}^*) \right\}, \end{aligned} \quad (15.19)$$

where $\widehat{\mathbf{p}}^{\text{ML}} = (\widehat{p}_1^{\text{ML}}, \dots, \widehat{p}_M^{\text{ML}})^\top = (\widehat{N}_1/N, \dots, \widehat{N}_M/N)^\top \in \Delta^{M-1}$ is the type, namely the empirical distribution of \mathcal{D} . Thus, if $\text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \widehat{\mathbf{p}}) > \text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \mathbf{p}^*)$, the right-hand side of Eq. (15.19) grows in the order of N . If $\mathcal{D} \in T_\varepsilon^N(\mathbf{p}^*)$, $\text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \mathbf{p}^*) = O_p(\varepsilon^2)$ since $\text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \mathbf{p}^*)$ is well approximated by a quadratic function of $\widehat{\mathbf{p}}^{\text{ML}} - \mathbf{p}^*$. To prove the first assertion of the lemma, it suffices to see that $\text{KL}(\mathbf{p}^* \parallel \widehat{\mathbf{p}}) \leq C\varepsilon^2$ is equivalent to $\text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \widehat{\mathbf{p}}) \leq C'\varepsilon^2$ for a constant $C' > 0$ if $\mathcal{D} \in T_\varepsilon^N(\mathbf{p}^*)$. In fact, we have

$$\text{KL}(\mathbf{p}^* \parallel \widehat{\mathbf{p}}) = \text{KL}(\mathbf{p}^* \parallel \widehat{\mathbf{p}}^{\text{ML}}) + \text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \widehat{\mathbf{p}}) + \sum_{m=1}^M (p_m^* - \widehat{p}_m^{\text{ML}}) (\log \widehat{p}_m^{\text{ML}} - \log \widehat{p}_m). \quad (15.20)$$

It follows from $\mathcal{D} \in T_\varepsilon^N(\mathbf{p}^*)$ that $\text{KL}(\mathbf{p}^* \parallel \widehat{\mathbf{p}}^{\text{ML}})/\varepsilon^2$ and $|p_m^* - \widehat{p}_m^{\text{ML}}|/\varepsilon$ are bounded by constants. Then $\text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \widehat{\mathbf{p}}) \leq C'\varepsilon^2$ implies that $|\widehat{p}_m^{\text{ML}} - \widehat{p}_m|/\varepsilon$ is bounded by a constant, and hence all the terms in Eq. (15.20) divided by ε^2 are bounded by constants.

It follows from Eq. (15.19) that

$$\min_{r_w(\mathbf{w})} \widetilde{Q} \geq -N \text{KL}(\widehat{\mathbf{p}}^{\text{ML}} \parallel \mathbf{p}^*). \quad (15.21)$$

The standard asymptotic theory of the multinomial model implies that twice the right-hand side of Eq. (15.21), with its sign flipped, asymptotically follows the chi-squared distribution with degree of freedom $M - 1$ as discussed in Section 13.4.5. \square

This lemma is used for proving the consistency of the VB posterior and evaluating lower-bounds of VB free energy for discrete models in Sections 15.3 and 15.4 and Chapter 16.

15.2 Mixture of Gaussians

In this section, we consider the following Gaussian mixture model (GMM) introduced in Section 4.1.1 and give upper- and lower-bounds for the VB free energy (Watanabe and Watanabe, 2004, 2006):

$$p(\mathbf{z}|\boldsymbol{\alpha}) = \text{Multinomial}_{K,1}(\mathbf{z}; \boldsymbol{\alpha}), \quad (15.22)$$

$$p(\mathbf{x}|\mathbf{z}, \{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{k=1}^K \{\text{Gauss}_M(\mathbf{x}; \boldsymbol{\mu}_k, \mathbf{I}_M)\}^{z_k}, \quad (15.23)$$

$$p(\boldsymbol{\alpha}|\boldsymbol{\phi}) = \text{Dirichlet}_K(\boldsymbol{\alpha}; (\boldsymbol{\phi}, \dots, \boldsymbol{\phi})^\top), \quad (15.24)$$

$$p(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, \xi) = \text{Gauss}_M(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, (1/\xi)\mathbf{I}_M). \quad (15.25)$$

Under the constraint,

$$r(\mathcal{H}, \mathbf{w}) = r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w}),$$

the VB posteriors are given as follows:

$$\begin{aligned} r(\{\mathbf{z}^{(n)}\}_{n=1}^N, \boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K) &= r_z(\{\mathbf{z}^{(n)}\}_{n=1}^N)r_\alpha(\boldsymbol{\alpha})r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K), \\ r_z(\{\mathbf{z}^{(n)}\}_{n=1}^N) &= \prod_{n=1}^N \text{Multinomial}_{K,1}\left(\mathbf{z}^{(n)}; \widehat{\mathbf{z}}^{(n)}\right), \\ r_\alpha(\boldsymbol{\alpha}) &= \text{Dirichlet}\left(\boldsymbol{\alpha}; (\widehat{\boldsymbol{\phi}}_1, \dots, \widehat{\boldsymbol{\phi}}_K)^\top\right), \\ r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K) &= \prod_{k=1}^K \text{Gauss}_M\left(\boldsymbol{\mu}_k; \widehat{\boldsymbol{\mu}}_k, \widehat{\sigma}_k^2 \mathbf{I}_M\right). \end{aligned}$$

The variational parameters $\{\mathbf{z}^{(n)}\}_{n=1}^N, \{\widehat{\boldsymbol{\phi}}_k\}_{k=1}^K, \{\widehat{\boldsymbol{\mu}}_k, \widehat{\sigma}_k^2\}_{k=1}^K$ minimize the free energy,

$$\begin{aligned} F &= \log\left(\frac{\Gamma(\sum_{k=1}^K \widehat{\boldsymbol{\phi}}_k)}{\prod_{k=1}^K \Gamma(\widehat{\boldsymbol{\phi}}_k)}\right) - \log\left(\frac{\Gamma(K\boldsymbol{\phi})}{(\Gamma(\boldsymbol{\phi}))^K}\right) - \frac{M}{2} \sum_{k=1}^K \log(\xi \widehat{\sigma}_k^2) - \frac{KM}{2} \\ &\quad + \sum_{n=1}^N \sum_{k=1}^K \widehat{z}_k^{(n)} \log \widehat{z}_k^{(n)} + \sum_{k=1}^K (\widehat{\boldsymbol{\phi}}_k - \boldsymbol{\phi} - \overline{\boldsymbol{\phi}}_k) (\Psi(\widehat{\boldsymbol{\phi}}_k) - \Psi(\sum_{k'=1}^K \widehat{\boldsymbol{\phi}}_{k'})) \end{aligned}$$

$$\begin{aligned}
& + \sum_{k=1}^K \frac{\xi (\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_0\|^2 + M\widehat{\sigma}_k^2)}{2} + \sum_{k=1}^K \frac{\overline{N}_k (M \log(2\pi) + M\widehat{\sigma}_k^2)}{2} \\
& + \sum_{k=1}^K \frac{\overline{N}_k \|\bar{\mathbf{x}}_k - \widehat{\boldsymbol{\mu}}_k\|^2 + \sum_{n=1}^N \widehat{z}_k^{(n)} \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}_k\|^2}{2}, \tag{15.26}
\end{aligned}$$

where

$$\overline{N}_k = \sum_{n=1}^N \widehat{z}_k^{(n)}, \tag{15.27}$$

$$\bar{\mathbf{x}}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^N \mathbf{x}^{(n)} \widehat{z}_k^{(n)}. \tag{15.28}$$

The stationary condition of the free energy yields

$$\widehat{z}_k^{(n)} = \frac{\bar{z}_k^{(n)}}{\sum_{k'=1}^K \bar{z}_{k'}^{(n)}}, \tag{15.29}$$

$$\widehat{\phi}_k = \overline{N}_k + \phi, \tag{15.30}$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{\overline{N}_k \bar{\mathbf{x}}_k + \xi \boldsymbol{\mu}_0}{\overline{N}_k + \xi}, \tag{15.31}$$

$$\widehat{\sigma}_k^2 = \frac{1}{\overline{N}_k + \xi}, \tag{15.32}$$

where

$$\bar{z}_k^{(n)} \propto \exp \left(\Psi(\widehat{\phi}_k) - \frac{1}{2} \|\mathbf{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\|^2 + M\widehat{\sigma}_k^2 \right). \tag{15.33}$$

The following condition is assumed.

Assumption 15.1 *The true distribution $q(\mathbf{x})$ is an M -dimensional GMM $p(\mathbf{x}|\mathbf{w}^*)$, which has K_0 components and parameter $\mathbf{w}^* = (\boldsymbol{\alpha}^*, \{\boldsymbol{\mu}_k^*\}_{k=1}^{K_0})$:*

$$q(\mathbf{x}) = p(\mathbf{x}|\mathbf{w}^*) = \sum_{k=1}^{K_0} \alpha_k^* \text{Gauss}_M(\mathbf{x}; \boldsymbol{\mu}_k^*, \mathbf{I}_M), \tag{15.34}$$

where $\mathbf{x}, \boldsymbol{\mu}_k^* \in \mathbb{R}^M$. Suppose that the true distribution can be realized by our model in hand, i.e., $K \geq K_0$ holds.

Under this condition, we prove the following theorem, which evaluates the relative VB free energy,

$$\widetilde{F}^{\text{VB}}(\mathcal{D}) = F^{\text{VB}}(\mathcal{D}) - NS_N(\mathcal{D}). \tag{15.35}$$

The proof will appear in the next section.

Theorem 15.5 *The relative VB free energy of the GMM satisfies*

$$\underline{\lambda}'^{\text{VB}}_{\text{MM}} \log N + NE_N(\widehat{\mathbf{w}}) + O_p(1) \leq \bar{F}^{\text{VB}}(\mathcal{D}) \leq \bar{\lambda}'^{\text{VB}}_{\text{MM}} \log N + O_p(1), \quad (15.36)$$

where E_N is the empirical KL divergence (15.14), and the coefficients $\underline{\lambda}'^{\text{VB}}_{\text{MM}}$, $\bar{\lambda}'^{\text{VB}}_{\text{MM}}$ are given by

$$\underline{\lambda}'^{\text{VB}}_{\text{MM}} = \begin{cases} (K-1)\phi + \frac{M}{2} & \left(\phi < \frac{M+1}{2}\right), \\ \frac{MK+K-1}{2} & \left(\phi \geq \frac{M+1}{2}\right), \end{cases} \quad (15.37)$$

$$\bar{\lambda}'^{\text{VB}}_{\text{MM}} = \begin{cases} (K-K_0)\phi + \frac{MK_0+K_0-1}{2} & \left(\phi < \frac{M+1}{2}\right), \\ \frac{MK+K-1}{2} & \left(\phi \geq \frac{M+1}{2}\right). \end{cases} \quad (15.38)$$

In this theorem, $E_N(\widehat{\mathbf{w}})$ is the training error of the VB estimator. Let $\widehat{\mathbf{w}}^{\text{ML}}$ be the ML estimator. Then it immediately follows from Eq. (15.14) that

$$NE_N(\widehat{\mathbf{w}}) \geq NE_N(\widehat{\mathbf{w}}^{\text{ML}}), \quad (15.39)$$

where $NE_N(\widehat{\mathbf{w}}^{\text{ML}}) = \min_{\mathbf{w}} \sum_{n=1}^N \log \frac{p(x^{(n)}|\mathbf{w}^*)}{p(x^{(n)}|\mathbf{w})}$ is the (maximum) log-likelihood ratio statistic with sign inversion. As discussed in Section 13.5.3, it is conjectured for the GMM defined by Eqs. (15.22) and (15.23) that the log-likelihood ratio diverges in the order of $\log \log N$ (Hartigan, 1985). If this conjecture is proved, the statement of the theorem is simplified to

$$\bar{F}^{\text{VB}}(\mathcal{D}) = \lambda' \log N + o_p(\log N),$$

for $\underline{\lambda}'^{\text{VB}}_{\text{MM}} \leq \lambda' \leq \bar{\lambda}'^{\text{VB}}_{\text{MM}}$. Note, however, that even if $NE_N(\widehat{\mathbf{w}}^{\text{ML}})$ diverges to minus infinity, Eq. (15.39) does not necessarily mean $NE_N(\widehat{\mathbf{w}})$ diverges in the same order. Also note that $NE_N(\widehat{\mathbf{w}})$ does not affect the upper-bound in Eq. (15.36).

Since the dimension of the parameter \mathbf{w} is $D = MK + K - 1$, the relative Bayes free energy coefficient of regular statistical models, on which the Bayesian information criterion (BIC) (Schwarz, 1978) and the minimum description length (MDL) (Rissanen, 1986) are based, is given by $D/2$. Note that, unlike regular models, the advantage of Bayesian learning for singular models is demonstrated by the asymptotic analysis as seen in Eqs. (13.123), (13.124), and (13.125). Theorem 15.5 claims that the coefficient $\bar{\lambda}'^{\text{VB}}_{\text{MM}}$ of $\log N$ is smaller than $D/2$ when $\phi < (M+1)/2$. This means that the VB free energy F^{VB} becomes smaller than that of regular models, i.e., $2\lambda'^{\text{VB}} \leq D$ holds.

Theorem 15.5 shows how the hyperparameters affect the learning process. The coefficients $\underline{\lambda}'^{\text{VB}}_{\text{MM}}$ and $\bar{\lambda}'^{\text{VB}}_{\text{MM}}$ in Eqs. (15.37) and (15.38) are divided into

two cases. These cases correspond to whether $\phi < \frac{M+1}{2}$ holds, indicating that the influence of the hyperparameter ϕ in the prior $p(\alpha|\phi)$ appears depending on the number M of parameters in each component. Let \widehat{K} be the number of components satisfying $\bar{N}_k = \Theta_p(N)$. Then the following corollary follows from the proof of Theorem 15.5.

Corollary 15.6 *The upper-bound in Eq. (15.36) is attained when $\widehat{K} = K_0$ if $\phi < \frac{M+1}{2}$ and $\widehat{K} = K$ if $\phi \geq \frac{M+1}{2}$.*

This corollary implies that the phase transition of the VB posterior occurs at $\phi = \frac{M+1}{2}$, i.e., only when $\phi < \frac{M+1}{2}$, the prior distribution reduces redundant components; otherwise, it uses all the components. The phase transition of the posterior occurs also in Bayesian learning while the phase transition point is different from that of VB learning (Yamazaki and Kaji, 2013).

Theorem 15.5 also implies that the hyperparameter ϕ is the only hyperparameter on which the leading term of the VB free energy F^{VB} depends. This is due to the influence of the hyperparameters on the prior probability density around the true parameters. Consider the case where $K_0 < K$. In this case, for a parameter that gives the true distribution, either of the followings holds: $\alpha_k = 0$ for some k or $\mu_i = \mu_j$ for some pair (i, j) . The prior distribution $p(\alpha|\phi)$ given by Eq. (15.24) can drastically change the probability density around the points where $\alpha_k = 0$ for some k by changing the hyperparameter ϕ while the prior distribution $p(\mu_k|\mu_0, \xi)$ given by Eq. (15.25) always takes positive values for any values of the hyperparameters ξ and μ_0 . While the condition for the prior density $p(\alpha|\phi)$ to diverge at $\alpha_k = 0$ is $\alpha_k < 1$, and hence is independent of M , the phase transition point of the VB posterior is $\phi = \frac{M+1}{2}$. As we will see in Section 15.4 for the Bernoulli mixture model, if some of the components are located at the boundary of the parameter space, the leading term of the relative VB free energy depends also on the hyperparameter of the prior for component parameters.

Theorem 15.5 is also extended to the case of the general Dirichlet prior $p(\alpha|\phi) = \text{Dirichlet}_K(\alpha; \phi)$, where $\phi = (\phi_1, \dots, \phi_K)^\top$ is the hyperparameter as follows:

Theorem 15.7 *(Nakamura and Watanabe, 2014) The relative VB free energy of the GMM satisfies*

$$\sum_{k=1}^K \lambda'_k \log N + NE_N(\widehat{w}) + O_p(1) \leq \widetilde{F}^{\text{VB}}(\mathcal{D}) \leq \sum_{k=1}^K \bar{\lambda}'_k \log N + O_p(1),$$

where the coefficients $\underline{\lambda}'^{\text{VB}}$, $\bar{\lambda}'^{\text{VB}}$ are given by

$$\begin{aligned}\underline{\lambda}'^{\text{VB}}_k &= \begin{cases} \phi_k - \frac{1}{2K} & (k \neq 1 \text{ and } \phi_k < \frac{M+1}{2}), \\ \frac{M+1}{2} - \frac{1}{2K} & (k = 1 \text{ or } \phi_k \geq \frac{M+1}{2}), \end{cases} \\ \bar{\lambda}'^{\text{VB}}_k &= \begin{cases} \phi_k - \frac{1}{2K} & (k > K_0 \text{ and } \phi_k < \frac{M+1}{2}), \\ \frac{M+1}{2} - \frac{1}{2K} & (k \leq K_0 \text{ or } \phi_k \geq \frac{M+1}{2}). \end{cases}\end{aligned}$$

The proof of this theorem is omitted. This theorem implies that the phase transition of the VB posterior of each component occurs at the same transition point $\phi_k = \frac{M+1}{2}$ as Theorem 15.5.

Proof of Theorem 15.5

Before proving Theorem 15.5, we show two lemmas where the two terms, $R = \text{KL}(r_w(\mathbf{w})||p(\mathbf{w}))$ and $Q = -\log C_{\mathcal{H}}$, in Lemma 15.1 are respectively evaluated.

Lemma 15.8 *It holds that*

$$\left| R - \left\{ G(\widehat{\boldsymbol{\alpha}}) + \frac{\xi}{2} \sum_{k=1}^K \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_0\|^2 \right\} \right| \leq C,$$

where C is a constant, $\widehat{\boldsymbol{\mu}}_k = \langle \boldsymbol{\mu}_k \rangle_{r_{\mu}(\{\boldsymbol{\mu}_k\}_{k=1}^K)} = \frac{\bar{N}_k \bar{\boldsymbol{\alpha}}_k + \xi \boldsymbol{\mu}_0}{\bar{N}_k + \xi}$, and the function $G(\widehat{\boldsymbol{\alpha}})$ of $\widehat{\boldsymbol{\alpha}} = \left\{ \widehat{\alpha}_k = \langle \alpha_k \rangle_{r_{\alpha}(\boldsymbol{\alpha})} = \frac{\bar{N}_k + \phi}{N + K\phi} \right\}_{k=1}^K$ is defined by

$$G(\widehat{\boldsymbol{\alpha}}) = \frac{MK + K - 1}{2} \log N + \left\{ \frac{M + 1}{2} - \phi \right\} \sum_{k=1}^K \log \widehat{\alpha}_k. \quad (15.40)$$

Proof Calculating the KL divergence between the posterior and the prior, we obtain

$$\text{KL}(r_{\alpha}(\boldsymbol{\alpha})||p(\boldsymbol{\alpha}|\phi)) = \sum_{k=1}^K h(\bar{N}_k) - N\Psi(N + K\phi) + \log \Gamma(N + K\phi) + \log \frac{\Gamma(\phi)^K}{\Gamma(K\phi)}, \quad (15.41)$$

where we use the notation $h(x) = x\Psi(x + \phi) - \log \Gamma(x + \phi)$. Similarly, we obtain

$$\begin{aligned}\text{KL}(r_{\mu}(\{\boldsymbol{\mu}_k\}_{k=1}^K)||p(\{\boldsymbol{\mu}_k\}_{k=1}^K|\boldsymbol{\mu}_0, \xi)) \\ = \sum_{k=1}^K \frac{M}{2} \log \frac{\bar{N}_k + \xi}{\xi} - \frac{KM}{2} + \frac{\xi}{2} \sum_{k=1}^K \left\{ \frac{M}{\bar{N}_k + \xi} + \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_0\|^2 \right\}.\end{aligned} \quad (15.42)$$

By using Inequalities (15.8) and (15.9), we obtain

$$-1 + \frac{12\phi - 1}{12(x + \phi)} \leq h(x) + \left(\phi - \frac{1}{2} \right) \log(x + \phi) - x - \phi + \frac{1}{2} \log 2\pi \leq 0. \quad (15.43)$$

Thus, from Eqs. (15.41), (15.42), (15.43), and

$$R = \text{KL}(r_\alpha(\boldsymbol{\alpha}) \| p(\boldsymbol{\alpha}|\phi)) + \text{KL}(r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K) \| p(\{\boldsymbol{\mu}_k\}_{k=1}^K | \boldsymbol{\mu}_0, \xi)),$$

it follows that

$$\begin{aligned} & \left| R - \left\{ G(\widehat{\boldsymbol{\alpha}}) + \frac{\xi}{2} \sum_{k=1}^K \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_0\|^2 \right\} \right| \\ & \leq \frac{MK + K - 1}{2} \log \left(1 + \frac{K\phi}{N} \right) \\ & \quad + (K-1) \left| \phi - \frac{\log 2\pi}{2} \right| + K + \sum_{k=1}^K \frac{|12\phi - 1|}{12(\bar{N}_k + \phi)} + \frac{12N + 1}{12(N + K\phi)} \\ & \quad + \left| \log \frac{\Gamma(\phi)^K}{\Gamma(K\phi)} \right| + \left| \sum_{k=1}^K \log \frac{\bar{N}_k + \xi}{\bar{N}_k + \phi} - \frac{MK}{2} (1 + \log \xi) + \frac{\xi}{2} \sum_{k=1}^K \frac{M}{\bar{N}_k + \xi} \right|. \end{aligned}$$

The right-hand side of the preceding inequality is bounded by a constant since

$$\frac{1}{N + \xi} < \frac{1}{\bar{N}_k + \xi} < \frac{1}{\xi},$$

and

$$\frac{1}{N + \phi} < \frac{1}{\bar{N}_k + \phi} < \frac{1}{\phi}.$$

□

Lemma 15.9 *It holds that*

$$\begin{aligned} Q = & - \sum_{n=1}^N \log \left(\sum_{k=1}^K \frac{1}{\sqrt{2\pi}^M} \exp \left(\Psi(\bar{N}_k + \phi) - \Psi(N + K\phi) \right. \right. \\ & \left. \left. - \frac{\|\mathbf{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\|^2}{2} - \frac{M}{2} \frac{1}{\bar{N}_k + \xi} \right) \right), \end{aligned} \quad (15.44)$$

and

$$NE_N(\widehat{\mathbf{w}}) - \frac{N}{N + K\phi} \leq \widetilde{Q} \leq N\overline{E}_N(\widehat{\mathbf{w}}) - \frac{N}{2(N + K\phi)}, \quad (15.45)$$

where $E_N(\widehat{\mathbf{w}})$ is given by Eq. (15.14) and $\overline{E}_N(\widehat{\mathbf{w}})$ is defined by

$$\overline{E}_N(\widehat{\mathbf{w}}) = \frac{1}{N} \sum_{n=1}^N \log \frac{p(\mathbf{x}^{(n)} | \mathbf{w}^*)}{\sum_{k=1}^K \frac{\widehat{\alpha}_k}{\sqrt{2\pi}^M} \exp \left(-\frac{\|\mathbf{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\|^2}{2} - \frac{M+2}{2(\bar{N}_k + \min\{\phi, \xi\})} \right)}.$$

Proof

$$\begin{aligned} C_{\mathcal{H}} &= \prod_{n=1}^N \sum_{z^{(n)}} \exp \left\langle \log p(\mathbf{x}^{(n)}, z^{(n)} | \mathbf{w}) \right\rangle_{r_w(\mathbf{w})} \\ &= \prod_{n=1}^N \sum_{k=1}^K \frac{1}{\sqrt{2\pi^M}} \exp \left(\Psi(\bar{N}_k + \phi) - \Psi(N + K\phi) \right. \\ &\quad \left. - \frac{\|\mathbf{x}^{(n)} - \hat{\boldsymbol{\mu}}_k\|^2}{2} - \frac{M}{2} \frac{1}{\bar{N}_k + \xi} \right). \end{aligned}$$

Thus, we have Eq. (15.44).

Using again Inequality (15.8), we obtain

$$\begin{aligned} Q &\leq - \sum_{n=1}^N \log \left(\sum_{k=1}^K \frac{\hat{\alpha}_k}{\sqrt{2\pi^M}} \exp \left(-\frac{\|\mathbf{x}^{(n)} - \hat{\boldsymbol{\mu}}_k\|^2}{2} - \frac{M+2}{2(\bar{N}_k + \min\{\phi, \xi\})} \right) \right) \\ &\quad - \frac{N}{2(N+K\phi)}, \end{aligned} \tag{15.46}$$

and

$$Q \geq - \sum_{n=1}^N \log \left(\sum_{k=1}^K \frac{\hat{\alpha}_k}{\sqrt{2\pi^M}} \exp \left(-\frac{\|\mathbf{x}^{(n)} - \hat{\boldsymbol{\mu}}_k\|^2}{2} \right) \right) - \frac{N}{N+K\phi},$$

which give upper- and lower-bounds in Eq. (15.45), respectively. \square

Now, from the preceding lemmas, we prove Theorem 15.5 by showing upper- and lower-bounds, respectively. First, we show the upper-bound in Eq. (15.36).

From Lemma 15.1, Lemma 15.8, and Lemma 15.9, it follows that

$$F - NS_N(\mathcal{D}) \leq \min_{\widehat{\mathbf{w}}} T_N(\widehat{\mathbf{w}}) + C, \tag{15.47}$$

where

$$T_N(\widehat{\mathbf{w}}) = G(\widehat{\boldsymbol{\alpha}}) + \frac{\xi}{2} \sum_{k=1}^K \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_0\|^2 + N\bar{E}_N(\widehat{\mathbf{w}}).$$

From Eq. (15.47), it is noted that the function values of $T_N(\widehat{\mathbf{w}})$ at specific points of the variational parameter $\widehat{\mathbf{w}}$ give upper-bounds of the VB free energy $F^{\text{VB}}(\mathcal{D})$. Hence, let us consider following two cases.

- (I) Consider the case where all components, including redundant ones, are used to learn K_0 true components, i.e.,

$$\widehat{\alpha}_k = \frac{\alpha_k^* N + \phi}{N + K\phi} \quad (1 \leq k \leq K_0 - 1),$$

$$\begin{aligned}\widehat{\alpha}_k &= \frac{\alpha_{K_0}^* N/(K - K_0 + 1) + \phi}{N + K\phi} && (K_0 \leq k \leq K), \\ \widehat{\mu}_k &= \mu_k^* && (1 \leq k \leq K_0 - 1), \\ \widehat{\mu}_k &= \mu_{K_0}^* && (K_0 \leq k \leq K).\end{aligned}$$

Then we obtain

$$\begin{aligned}& N\bar{E}_N(\widehat{\mathbf{w}}) \\ & < \sum_{n=1}^N \log p(\mathbf{x}^{(n)}|\mathbf{w}^*) - \sum_{n=1}^N \log \frac{N + \phi}{N + K\phi} \\ & \quad - \sum_{n=1}^N \log \left(\sum_{k=1}^{K_0-1} \frac{\alpha_k^*}{\sqrt{2\pi}^M} \exp \left(-\frac{\|\mathbf{x}^{(n)} - \mu_k^*\|^2}{2} - \frac{M+2}{2(\alpha_k^* N + \min\{\xi, \phi\})} \right) \right. \\ & \quad \left. + \frac{\alpha_{K_0}^*}{\sqrt{2\pi}^M} \exp \left(-\frac{\|\mathbf{x}^{(n)} - \mu_{K_0}^*\|^2}{2} - \frac{M+2}{2(\frac{\alpha_{K_0}^*}{K-K_0+1} N + \min\{\xi, \phi\})} \right) \right) \\ & < \sum_{n=1}^N \log \frac{\frac{N+K\phi}{N+\phi} p(\mathbf{x}^{(n)}|\mathbf{w}^*)}{p(\mathbf{x}^{(n)}|\mathbf{w}^*) \exp \left(-\frac{(M+2)(K-K_0+1)}{2(\min_k \{\alpha_k^*\} N + \min\{\xi, \phi\}(K-K_0+1))} \right)} \\ & < \frac{(K-1)\phi}{N+\phi} + \frac{(M+2)(K-K_0+1)N}{2(\min_k \{\alpha_k^*\} N + \min\{\xi, \phi\}(K-K_0+1))} \\ & \leq (K-1)\phi + \left(\frac{M+2}{2} \right) \frac{K-K_0+1}{\min_k \{\alpha_k^*\}},\end{aligned}$$

where the first inequality follows from $\frac{\alpha_k^* N + \phi}{N + K\phi} > \alpha_k^* \frac{N + \phi}{N + K\phi}$ and the third inequality follows from $\log(1+x) \leq x$ for $x > -1$.

It follows that

$$T_N(\widehat{\mathbf{w}}) < \frac{MK + K - 1}{2} \log N + C', \quad (15.48)$$

where C' is a constant.

(II) Consider the case where the redundant components are eliminated, i.e.,

$$\begin{aligned}\widehat{\alpha}_k &= \frac{\alpha_k^* N + \phi}{N + K\phi} && (1 \leq k \leq K_0), \\ \widehat{\alpha}_k &= \frac{\phi}{N + K\phi} && (K_0 + 1 \leq k \leq K), \\ \widehat{\mu}_k &= \mu_k^* && (1 \leq k \leq K_0), \\ \widehat{\mu}_k &= \mu_0 && (K_0 + 1 \leq k \leq K).\end{aligned}$$

Then it holds that

$$\begin{aligned}
& N\bar{E}_N(\widehat{\boldsymbol{w}}) \\
& < \sum_{n=1}^N \log \frac{p(\mathbf{x}^{(n)}|\boldsymbol{w}^*)}{\frac{N+\phi}{N+K\phi} \sum_{k=1}^{K_0} \frac{\alpha_k^*}{\sqrt{2\pi^M}} \exp\left(-\frac{\|\mathbf{x}^{(n)} - \boldsymbol{\mu}_k^*\|^2}{2} - \frac{M+2}{2(\alpha_k^* N + \min\{\xi, \phi\})}\right)} \\
& < \frac{(K-1)\phi N}{N+\phi} + \left(\frac{M+2}{2}\right) \frac{N}{\min_k\{\alpha_k^*\} N + \min\{\xi, \phi\}} \\
& \leq (K-1)\phi + \left(\frac{M+2}{2}\right) \frac{1}{\min_k\{\alpha_k^*\}}. \tag{15.49}
\end{aligned}$$

The first inequality follows from $\frac{\alpha_k^* N + \phi}{N + K\phi} > \alpha_k^* \frac{N + \phi}{N + K\phi}$ and

$$\sum_{k=K_0+1}^K \frac{\widehat{\alpha}_k}{\sqrt{2\pi^M}} \exp\left(-\frac{\|\mathbf{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\|^2}{2} - \frac{M+2}{2(\bar{N}_k + \min\{\phi, \xi\})}\right) > 0.$$

The second inequality follows from $\log(1+x) \leq x$ for $x > -1$.

It follows that

$$T_N(\widehat{\boldsymbol{w}}) < \left\{ (K-K_0)\phi + \frac{MK_0 + K_0 - 1}{2} \right\} \log N + C'', \tag{15.50}$$

where C'' is a constant.

From Eqs. (15.47), (15.48), and (15.50), we obtain the upper-bound in Eq. (15.36).

Next we show the lower-bound in Eq. (15.36). It follows from Lemma 15.1, Lemma 15.8, and Lemma 15.9 that

$$F - NS_N(\mathcal{D}) \geq \min_{\widehat{\boldsymbol{\alpha}}} \{G(\widehat{\boldsymbol{\alpha}})\} + NE_N(\widehat{\boldsymbol{w}}) - C - 1. \tag{15.51}$$

If $\phi \geq \frac{M+1}{2}$, then,

$$G(\widehat{\boldsymbol{\alpha}}) \geq \frac{MK + K - 1}{2} \log N - \left(\frac{M+1}{2} - \phi\right) K \log K, \tag{15.52}$$

since Jensen's inequality yields that

$$\sum_{k=1}^K \log \widehat{\alpha}_k \leq K \log \left(\frac{1}{K} \sum_{k=1}^K \widehat{\alpha}_k \right) = K \log \left(\frac{1}{K} \right).$$

If $\phi < \frac{M+1}{2}$, then

$$\begin{aligned} G(\widehat{\alpha}) &\geq \left\{ (K-1)\phi + \frac{M}{2} \right\} \log N + \left(\frac{M+1}{2} - \phi \right) (K-1) \log \frac{\phi N}{N+K\phi} + C''' \\ &\geq \left\{ (K-1)\phi + \frac{M}{2} \right\} \log N + \left(\frac{M+1}{2} - \phi \right) (K-1) \log \frac{\phi}{1+K\phi} + C''', \end{aligned} \quad (15.53)$$

where C''' is a constant. The first inequality follows since

$$\widehat{\alpha}_k \geq \frac{\phi}{N+K\phi}$$

holds for every k , and the constraint

$$\sum_{k=1}^K \widehat{\alpha}_k = 1$$

ensures that $|\log \widehat{\alpha}_k|$ is bounded by a constant independent of N for at least one index k . From Eqs. (15.51), (15.52), and (15.53) we obtain the lower-bound in Eq. (15.36).

15.3 Mixture of Exponential Family Distributions

The previous theorem for the GMM can be generalized to the mixture of *exponential family distributions* (Watanabe and Watanabe, 2005, 2007). The model that we consider is defined by

$$p(z|\alpha) = \text{Multinomial}_{K,1}(z; \alpha), \quad (15.54)$$

$$p(t|z, \{\eta_k\}_{k=1}^K) = \prod_{k=1}^K \left\{ \exp \left(\eta_k^\top t - A(\eta_k) + B(t) \right) \right\}^{z_k}, \quad (15.55)$$

$$p(\alpha|\phi) = \text{Dirichlet}_K(\alpha; (\phi, \dots, \phi)^\top), \quad (15.56)$$

$$p(\eta_k|\nu_0, \xi) = \frac{1}{C(\xi, \nu_0)} \exp \left(\xi(\nu_0^\top \eta_k - A(\eta_k)) \right). \quad (15.57)$$

As demonstrated in Section 4.1.2, under the constraint, $r(\mathcal{H}, w) = r_{\mathcal{H}}(\mathcal{H})r_w(w)$, the VB posteriors are given as follows:

$$r(\{z^{(n)}\}_{n=1}^N, \alpha, \{\eta_k\}_{k=1}^K) = r_z(\{z^{(n)}\}_{n=1}^N) r_\alpha(\alpha) r_\eta(\{\eta_k\}_{k=1}^K),$$

$$r_z(\{z^{(n)}\}_{n=1}^N) = \prod_{n=1}^N \text{Multinomial}_{K,1}(z^{(n)}; \bar{z}^{(n)}),$$

$$r_\alpha(\boldsymbol{\alpha}) = \text{Dirichlet} \left(\boldsymbol{\alpha}; (\widehat{\phi}_1, \dots, \widehat{\phi}_K)^\top \right),$$

$$r_\eta(\{\boldsymbol{\eta}_k\}_{k=1}^K) = \prod_{k=1}^K \frac{1}{C(\widehat{\xi}_k, \widehat{\nu}_k)} \exp \left(\widehat{\xi}_k (\widehat{\nu}_k^\top \boldsymbol{\eta}_k - A(\boldsymbol{\eta}_k)) \right), \quad (15.58)$$

The variational parameters $\{\mathbf{z}^{(n)}\}_{n=1}^N, \{\widehat{\phi}_k\}_{k=1}^K, \{\widehat{\nu}_k, \widehat{\xi}_k\}_{k=1}^K$ minimize the free energy,

$$\begin{aligned} F &= \log \left(\frac{\Gamma(\sum_{k=1}^K \widehat{\phi}_k)}{\prod_{k=1}^K \Gamma(\widehat{\phi}_k)} \right) - \log \left(\frac{\Gamma(K\phi)}{(\Gamma(\phi))^K} \right) - \sum_{k=1}^K \log C(\widehat{\xi}_k, \widehat{\nu}_k) + K \log C(\xi, \nu_0) \\ &\quad + \sum_{n=1}^N \sum_{k=1}^K \widehat{z}_k^{(n)} \log \widehat{z}_k^{(n)} + \sum_{k=1}^K (\widehat{\phi}_k - \phi - \bar{N}_k) (\Psi(\widehat{\phi}_k) - \Psi(\sum_{k'=1}^K \widehat{\phi}_{k'})) \\ &\quad + \sum_{k=1}^K \left[\widehat{\boldsymbol{\eta}}_k^\top \left\{ \xi (\widehat{\nu}_k - \nu_0) + \bar{N}_k (\widehat{\nu}_k - \bar{\mathbf{t}}_k) \right\} + (\widehat{\xi}_k - \xi - \bar{N}_k) \frac{\partial \log C(\widehat{\xi}_k, \widehat{\nu}_k)}{\partial \xi_k} \right] \\ &\quad - \sum_{n=1}^N B(\mathbf{t}^{(n)}), \end{aligned} \quad (15.59)$$

where

$$\begin{aligned} \bar{N}_k &= \sum_{n=1}^N \widehat{z}_k^{(n)}, \\ \bar{\mathbf{t}}_k &= \frac{1}{\bar{N}_k} \sum_{n=1}^N \langle \widehat{z}_k^{(n)} \rangle_{r_{\mathcal{H}}(\mathcal{H})} \mathbf{t}^{(n)}, \\ \widehat{\boldsymbol{\eta}}_k &= \frac{1}{\widehat{\xi}_k} \frac{\partial \log C(\widehat{\xi}_k, \widehat{\nu}_k)}{\partial \nu_k}. \end{aligned}$$

The stationary condition of the free energy yields

$$\widehat{z}_k^{(n)} = \frac{\bar{z}_k^{(n)}}{\sum_{k'=1}^K \bar{z}_{k'}^{(n)}}, \quad (15.60)$$

$$\widehat{\alpha}_k = \bar{N}_k + \phi, \quad (15.61)$$

$$\widehat{\nu}_k = \frac{\bar{N}_k \bar{\mathbf{t}}_k + \xi \nu_0}{\bar{N}_k + \xi}, \quad (15.62)$$

$$\widehat{\xi}_k = \bar{N}_k + \xi. \quad (15.63)$$

where

$$\bar{z}_k^{(n)} \propto \exp \left(\Psi(\widehat{\phi}_k) + \widehat{\boldsymbol{\eta}}_k^\top \mathbf{t}^{(n)} - \langle A(\boldsymbol{\eta}_k) \rangle_{r_\eta(\boldsymbol{\eta}_k)} \right). \quad (15.64)$$

We assume the following conditions.

Assumption 15.2 *The true distribution $q(\mathbf{t})$ of sufficient statistics is represented by a mixture of exponential family distributions $p(\mathbf{t}|\mathbf{w}^*)$, which has K_0 components and the parameter $\mathbf{w}^* = \{\alpha_k^*, \boldsymbol{\eta}_k^*\}_{k=1}^{K_0}$:*

$$q(\mathbf{t}) = p(\mathbf{t}|\mathbf{w}^*) = \sum_{k=1}^{K_0} \alpha_k^* \exp\left(\boldsymbol{\eta}_k^{*\top} \mathbf{t} - A(\boldsymbol{\eta}_k^*) + B(\mathbf{t})\right),$$

where $\boldsymbol{\eta}_k^* \in \mathbb{R}^M$ and $\boldsymbol{\eta}_k^* \neq \boldsymbol{\eta}_{k'}^*$ ($k \neq k'$). Also, assume that the true distribution can be achieved with the model, i.e., $K \geq K_0$ holds.

Assumption 15.3 *The prior distribution $p(\{\boldsymbol{\eta}_k\}_{k=1}^K | \nu_0, \xi)$ defined by Eq. (15.57) satisfies $0 < p(\{\boldsymbol{\eta}_k\}_{k=1}^K | \nu_0, \xi) < \infty$.*

Assumption 15.4 *Regarding the distribution $p(\mathbf{t}|\boldsymbol{\eta})$ of each component, the Fisher information matrix*

$$\mathbf{F}(\boldsymbol{\eta}) = \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top}$$

satisfies $0 < \det(\mathbf{F}(\boldsymbol{\eta})) < +\infty$ for an arbitrary $\boldsymbol{\eta} \in \mathbf{H}$. The function $\boldsymbol{\nu}^\top \boldsymbol{\eta} - A(\boldsymbol{\eta})$ has a stationary point at $\widehat{\boldsymbol{\eta}}$ in the interior of \mathbf{H} for each $\boldsymbol{\nu} \in \left\{ \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} ; \boldsymbol{\eta} \in \mathbf{H} \right\}$.

The following theorem will be proven under these conditions. The proof will appear in the next section. Here,

$$S_N(\mathcal{D}) = -\frac{1}{N} \sum_{n=1}^N \log p(\mathbf{t}^{(n)} | \mathbf{w}^*) \quad (15.65)$$

is the empirical entropy.

Theorem 15.10 *The relative VB free energy of the mixture of exponential family distributions satisfies*

$$\begin{aligned} \underline{\lambda}_{\text{MM}}^{\text{VB}} \log N + N E_N(\widehat{\mathbf{w}}) + O_p(1) &\leq \widetilde{F}^{\text{VB}}(\mathcal{D}) = F^{\text{VB}}(\mathcal{D}) - N S_N(\mathcal{D}) \\ &\leq \overline{\lambda}_{\text{MM}}^{\text{VB}} \log N + O_p(1), \end{aligned} \quad (15.66)$$

where $\underline{\lambda}_{\text{MM}}^{\text{VB}}$ and $\overline{\lambda}_{\text{MM}}^{\text{VB}}$ are given by

$$\underline{\lambda}_{\text{MM}}^{\text{VB}} = \begin{cases} (K-1)\phi + \frac{M}{2} & (\phi < \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi \geq \frac{M+1}{2}), \end{cases} \quad (15.67)$$

$$\overline{\lambda}_{\text{MM}}^{\text{VB}} = \begin{cases} (K-K_0)\phi + \frac{MK_0+K_0-1}{2} & (\phi < \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi \geq \frac{M+1}{2}). \end{cases} \quad (15.68)$$

Again in this theorem,

$$E_N(\widehat{\mathbf{w}}) = \frac{1}{N} \sum_{n=1}^N \log \frac{p(\mathbf{t}^{(n)}|\mathbf{w}^*)}{p(\mathbf{t}^{(n)}|\widehat{\mathbf{w}})} \quad (15.69)$$

is the training error, and

$$NE_N(\widehat{\mathbf{w}}) \geq NE_N(\widehat{\mathbf{w}}^{\text{ML}}), \quad (15.70)$$

holds for the (maximum) log-likelihood ratio statistic. As discussed in Section 13.5.3, the log-likelihood ratio statistics of some singular models diverge to infinity as N increases. Some known facts about the divergence of the log-likelihood ratio are described in the following examples. Note again that even if $NE_N(\widehat{\mathbf{w}}^{\text{ML}})$ diverges to minus infinity, Eq. (15.70) does not necessarily mean $NE_N(\widehat{\mathbf{w}})$ diverges in the same order.

If the domain of the sufficient statistics \mathbf{t} of the model $p(\mathbf{t}|\mathbf{w})$ is discrete and finite, we obtain the following theorem by Lemmas 15.3 and 15.4:

Theorem 15.11 *If the domain of the sufficient statistics \mathbf{t} is discrete and finite, the relative VB free energy of the mixture of exponential family distributions satisfies*

$$\widetilde{F}^{\text{VB}}(\mathcal{D}) = \bar{\lambda}_{\text{MM}}^{\text{VB}} \log N + O_p(1), \quad (15.71)$$

where the coefficient $\bar{\lambda}_{\text{MM}}^{\text{VB}}$ is given by Eq. (15.68).

The proof of this theorem follows the proof of the preceding theorem.

Examples

The following are examples where Theorems 15.10 and 15.11 apply.

Example 1 (Binomial) Consider a mixture of binomial component distributions. Each component has a one-dimensional parameter $v \in [0, 1]$:

$$p(x = k|v) = \text{Binomial}_T(k; v) = \binom{T}{k} v^k (1-v)^{T-k}, \quad (15.72)$$

where T is the number of Bernoulli trials and $k = 0, 1, 2, \dots, T$. Hence, $M = 1$ and the natural parameter is given by $\eta = \log \frac{v}{1-v}$. Theorem 15.11 applies with $M = 1$.

Example 2 (Gamma) Consider the gamma component with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$:

$$p(x|\alpha, \beta) = \text{Gamma}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \quad (15.73)$$

where $0 \leq x < \infty$. The natural parameter $\boldsymbol{\eta}$ is given by $\eta_1 = \beta$ and $\eta_2 = \alpha - 1$. Hence, Eq. (15.66) holds where $\underline{\lambda}_{\text{MM}}^{\text{VB}}$ and $\bar{\lambda}_{\text{MM}}^{\text{VB}}$ are given by Eqs. (15.67) and (15.68) with $M = 2$. When shape parameter α is known, the likelihood ratio in ML learning diverges in the order of $\log \log N$ (Liu et al., 2003). This implies that $NE_N(\bar{\mathbf{w}}) = O_p(\log \log N)$ from Eq. (15.70).

Example 3 (Gaussian) Consider the L -dimensional Gaussian component with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{Gauss}_L(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{L/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right).$$

The natural parameter $\boldsymbol{\eta}$ is given by $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}^{-1}$. These are functions of the elements of $\boldsymbol{\mu}$ and the upper-right half of $\boldsymbol{\Sigma}^{-1}$. Hence, Eq. (15.66) holds where $\underline{\lambda}_{\text{MM}}^{\text{VB}}$ and $\bar{\lambda}_{\text{MM}}^{\text{VB}}$ are given by Eqs. (15.67) and (15.68) with $M = L + L(L+1)/2$. If the covariance matrix $\boldsymbol{\Sigma}$ is known and the parameter is restricted to mean $\boldsymbol{\mu}$, it is conjectured that the likelihood ratio in ML learning diverges in the order of $\log \log N$ (Hartigan, 1985). This suggests that the likelihood ratio can diverge in a higher order than $\log \log N$ if the covariance matrices are also estimated.

Other than these examples, Theorems 15.10 and 15.11 apply to mixtures of distributions such as multinomial, Poisson, and Weibull.

Proof of Theorem 15.10

Here Theorem 15.10 is proved in the same way as Theorem 15.5.

Since the VB posterior satisfies $r_w(\mathbf{w}) = r_\alpha(\boldsymbol{\alpha})r_\eta(\{\boldsymbol{\eta}_k\}_{k=1}^K)$, we have

$$\begin{aligned} R &= \text{KL}(r_w(\mathbf{w}) \| p(\mathbf{w})) \\ &= \text{KL}(r_\alpha(\boldsymbol{\alpha}) \| p(\boldsymbol{\alpha}|\phi)) + \sum_{k=1}^K \text{KL}(r_\eta(\boldsymbol{\eta}_k) \| p(\boldsymbol{\eta}_k|\boldsymbol{\nu}_0, \xi)). \end{aligned} \quad (15.74)$$

The following lemma is used for evaluating $\text{KL}(r_\eta(\boldsymbol{\eta}_k) \| p(\boldsymbol{\eta}_k|\boldsymbol{\nu}_0, \xi))$ in the case of the mixture of exponential family distributions.

Lemma 15.12 *It holds that*

$$\text{KL}(r_\eta(\boldsymbol{\eta}_k) \| p(\boldsymbol{\eta}_k|\boldsymbol{\nu}_0, \xi)) = \frac{M}{2} \log(\bar{N}_k + \xi) - \log p(\widehat{\boldsymbol{\eta}}_k|\boldsymbol{\nu}_0, \xi) + O_p(1),$$

where

$$\widehat{\boldsymbol{\eta}}_k = \langle \boldsymbol{\eta}_k \rangle_{r_\eta(\boldsymbol{\eta}_k)} = \frac{1}{\widehat{\xi}_k} \frac{\partial \log C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k)}{\partial \widehat{\boldsymbol{v}}_k}. \quad (15.75)$$

Proof Using the VB posterior, Eq. (15.58), we obtain

$$\text{KL}(r_\eta(\boldsymbol{\eta}_k) \| p(\boldsymbol{\eta}_k | \boldsymbol{v}_0, \xi)) = -\log \frac{C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k)}{C(\xi, \boldsymbol{v}_0)} + \overline{N}_k \left\{ \widehat{\boldsymbol{v}}_k \langle \boldsymbol{\eta}_k \rangle_{r_\eta(\boldsymbol{\eta}_k)} - \langle A(\boldsymbol{\eta}_k) \rangle_{r_\eta(\boldsymbol{\eta}_k)} \right\}, \quad (15.76)$$

where we used $\widehat{\xi}_k = \overline{N}_k + \xi$. Let us now evaluate the value of $C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k)$ when $\widehat{\xi}_k$ is sufficiently large. From Assumption 15.4, using the saddle point approximation, we obtain

$$C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k) = \exp \left(\widehat{\xi}_k \{ \widehat{\boldsymbol{v}}_k^\top \widetilde{\boldsymbol{\eta}}_k - A(\widetilde{\boldsymbol{\eta}}_k) \} \right) \left(\frac{2\pi}{\widehat{\xi}_k} \right)^{M/2} \sqrt{\det(\mathbf{F}(\widetilde{\boldsymbol{\eta}}_k))^{-1}} \left\{ 1 + O_p \left(\frac{1}{\widehat{\xi}_k} \right) \right\}, \quad (15.77)$$

where $\widetilde{\boldsymbol{\eta}}_k$ is the maximizer of the function $\widehat{\boldsymbol{v}}^\top \boldsymbol{\eta}_k - A(\boldsymbol{\eta}_k)$, that is,

$$\frac{\partial A(\widetilde{\boldsymbol{\eta}}_k)}{\partial \boldsymbol{\eta}_k} = \widehat{\boldsymbol{v}}_k.$$

Therefore, $-\log C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k)$ is evaluated as

$$-\log C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k) = \frac{M}{2} \log \frac{\widehat{\xi}_k}{2\pi} + \frac{1}{2} \log \det(\mathbf{F}(\widetilde{\boldsymbol{\eta}}_k)) - \widehat{\xi}_k \left\{ \widehat{\boldsymbol{v}}_k^\top \widetilde{\boldsymbol{\eta}}_k - A(\widetilde{\boldsymbol{\eta}}_k) \right\} + O_p \left(\frac{1}{\widehat{\xi}_k} \right). \quad (15.78)$$

Applying the saddle point approximation to

$$\boldsymbol{\eta}_k - \widetilde{\boldsymbol{\eta}}_k = \frac{1}{C(\widehat{\xi}_k, \widehat{\boldsymbol{v}}_k)} \int (\boldsymbol{\eta}_k - \widetilde{\boldsymbol{\eta}}_k) \exp \left(\widehat{\xi}_k \left\{ \widehat{\boldsymbol{v}}_k^\top \boldsymbol{\eta}_k - A(\boldsymbol{\eta}_k) \right\} \right) d\boldsymbol{\eta}_k,$$

we obtain

$$\|\widehat{\boldsymbol{\eta}}_k - \widetilde{\boldsymbol{\eta}}_k\| \leq \frac{A'}{\widehat{\xi}_k} + O_p \left(\widehat{\xi}_k^{-3/2} \right), \quad (15.79)$$

where A' is a constant. Since

$$A(\boldsymbol{\eta}_k) - A(\widetilde{\boldsymbol{\eta}}_k) = (\boldsymbol{\eta}_k - \widetilde{\boldsymbol{\eta}}_k)^\top \widehat{\boldsymbol{v}}_k + \frac{1}{2} (\boldsymbol{\eta}_k - \widetilde{\boldsymbol{\eta}}_k)^\top \mathbf{F}(\widetilde{\boldsymbol{\eta}}_k) (\boldsymbol{\eta}_k - \widetilde{\boldsymbol{\eta}}_k), \quad (15.80)$$

for some point $\bar{\boldsymbol{\eta}}_k$ on the line segment between $\boldsymbol{\eta}_k$ and $\tilde{\boldsymbol{\eta}}_k$, we have

$$A(\widehat{\boldsymbol{\eta}}_k) - A(\tilde{\boldsymbol{\eta}}_k) = (\widehat{\boldsymbol{\eta}}_k - \tilde{\boldsymbol{\eta}}_k)^\top \widehat{\mathbf{v}}_k + O_p(\widehat{\xi}_k^{-2}), \quad (15.81)$$

and applying the saddle point approximation, we obtain

$$\langle A(\boldsymbol{\eta}_k) \rangle_{r_\eta(\boldsymbol{\eta}_k)} - A(\tilde{\boldsymbol{\eta}}_k) = (\widehat{\boldsymbol{\eta}}_k - \tilde{\boldsymbol{\eta}}_k)^\top \widehat{\mathbf{v}}_k + \frac{M}{2\widehat{\xi}_k} + O_p(\widehat{\xi}_k^{-3/2}). \quad (15.82)$$

From Eqs. (15.81) and (15.82), we have

$$\langle A(\boldsymbol{\eta}_k) \rangle_{r_\eta(\boldsymbol{\eta}_k)} - A(\bar{\boldsymbol{\eta}}_k) = \frac{M}{2\widehat{\xi}_k} + O_p(\widehat{\xi}_k^{-3/2}), \quad (15.83)$$

Thus, from Eqs. (15.76), (15.78), (15.81), and (15.82), we obtain the lemma. \square

Lemmas 15.8 and 15.9 are substituted by the following lemmas.

Lemma 15.13 *It holds that*

$$\left| R - G(\widehat{\alpha}) + \sum_{k=1}^K \log p(\widehat{\boldsymbol{\eta}}_k | \mathbf{v}_0, \xi) \right| \leq C, \quad (15.84)$$

where C is a constant and the function $G(\widehat{\alpha})$ is defined by Eq. (15.40).

Proof From Eqs. (15.41), (15.43), and (15.74) and Lemma 15.12,

$$\left| R - G(\widehat{\alpha}) + \sum_{k=1}^K \log p(\widehat{\boldsymbol{\eta}}_k | \mathbf{v}_0, \xi) \right|$$

is upper-bounded by a constant since

$$\frac{1}{N + \xi} < \frac{1}{\bar{N}_k + \xi} < \frac{1}{\xi}.$$

\square

Lemma 15.14 *It holds that*

$$NE_N(\widehat{\mathbf{w}}) + O_p(1) \leq \widetilde{Q} = -\log C_{\mathcal{H}} - NS_N(\mathcal{D}) \leq N\overline{E}_N(\widehat{\mathbf{w}}) + O_p(1), \quad (15.85)$$

where the function $E_N(\mathbf{w})$ is defined by Eq. (15.69) and

$$\overline{E}_N(\widehat{\mathbf{w}}) = \frac{1}{N} \sum_{n=1}^N \log \frac{p(\mathbf{t}^{(n)} | \mathbf{w}^*)}{\sum_{k=1}^K \widehat{\alpha}_k p(\mathbf{t}^{(n)} | \widehat{\boldsymbol{\eta}}_k) \exp\left(-\frac{A}{\bar{N}_k + \min\{\phi, \xi\}}\right)},$$

where A is a constant.

Proof

$$\begin{aligned} C_{\mathcal{H}} &= \prod_{n=1}^N \sum_{k=1}^K \exp \left\langle \log \alpha_k p(\boldsymbol{t}^{(n)} | \boldsymbol{\eta}_k) \right\rangle_{r_w(\boldsymbol{w})} \\ &= \prod_{n=1}^N \sum_{k=1}^K \exp \left(\Psi(\bar{N}_k + \phi) - \Psi(N + K\phi) + \widehat{\boldsymbol{\eta}}_k^\top \boldsymbol{t}^{(n)} - \langle A(\boldsymbol{\eta}_k) \rangle_{r_\eta(\boldsymbol{\eta}_k)} + B(\boldsymbol{t}^{(n)}) \right). \end{aligned}$$

Again, using the inequalities in Eqs. (15.8) and (15.83), we obtain

$$\begin{aligned} Q &\leq \sum_{n=1}^N \log \left(\sum_{k=1}^K \widehat{\alpha}_k p(\boldsymbol{t}^{(n)} | \widehat{\boldsymbol{\eta}}_k) \exp \left(-\frac{M+2}{2(\bar{N}_k + \min\{\phi, \xi\})} + O_p\left(\bar{N}_k^{-\frac{3}{2}}\right) \right) \right) + O_p(1), \\ Q &\geq - \sum_{n=1}^N \log \left(\sum_{k=1}^K \widehat{\alpha}_k p(\boldsymbol{t}^{(n)} | \widehat{\boldsymbol{\eta}}_k) \right) + O_p(1), \end{aligned}$$

which give the upper- and lower-bounds in Eq. (15.85), respectively. \square

Since the prior distribution $p(\{\boldsymbol{\eta}_k\}_{k=1}^K | \boldsymbol{v}_0, \xi)$ satisfies $0 < p(\{\boldsymbol{\eta}_k\}_{k=1}^K | \boldsymbol{v}_0, \xi) < \infty$, from Lemmas 15.13 and 15.14, we complete the proof of Theorem 15.10 in the same way as that of Theorem 15.5.

Proof of Theorem 15.11

The upper-bound follows from Theorem 15.10. From Lemmas 15.1 and 15.13 and the boundedness of the prior, we have the following lower-bound:

$$F - NS_N(\mathcal{D}) \geq G(\widehat{\boldsymbol{\alpha}}) + \widetilde{Q} + O_p(1).$$

Lemma 15.4 implies that for $\varepsilon > 0$, if $\mathcal{D} \in T_\varepsilon^N(\boldsymbol{p}^*)$ and $\widehat{\boldsymbol{p}} \notin R_{C\varepsilon^2}^*$ for the constant C in the lemma,

$$\widetilde{Q} = \Omega_p(N).$$

Since $G(\widehat{\boldsymbol{\alpha}}) = O_p(\log N)$, this means that if the free energy is minimized, $\widehat{\boldsymbol{p}} \in R_{C\varepsilon^2}^*$ for sufficiently large N , which implies that at least K_0 components are active and

$$|\log \widehat{\alpha}_k| = O_p(1)$$

holds for at least K_0 components. By minimizing $G(\widehat{\boldsymbol{\alpha}})$ under this constraint and the second assertion of Lemma 15.4, we have

$$F^{\text{VB}}(\mathcal{D}) - NS_N(\mathcal{D}) \geq \overline{\lambda}_{\text{MM}}^{\text{VB}} \log N + O_p(1),$$

for $\mathcal{D} \in T_\varepsilon^N(\boldsymbol{p}^*)$. Because the probability that the observed data sequence is strongly ε -typical tends to 1 as $N \rightarrow \infty$ for any $\varepsilon > 0$ by Lemma 15.3, we obtain the theorem.

15.4 Mixture of Bernoulli with Deterministic Components

In the previous sections, we assumed that all true component parameters are in the interior of the parameter space. In this section, we consider the *Bernoulli mixture model* when some components are at the boundary of the parameter space (Kaji et al., 2010).

For an M -dimensional binary vector, $\mathbf{x} = (x_1, \dots, x_M)^\top \in \{0, 1\}^M$, we define the Bernoulli distribution with parameter $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)^\top$ as

$$\text{Bern}_M(\mathbf{x}|\boldsymbol{\mu}) = \prod_{m=1}^M \mu_m^{x_m} (1 - \mu_m)^{(1-x_m)}.$$

For each element of $\boldsymbol{\mu}$, its conjugate prior, the Beta distribution, is given by

$$\text{Beta}(\mu; a, b) = \frac{1}{\mathcal{B}(a, b)} \mu^{a-1} (1 - \mu)^{b-1},$$

for $a, b > 0$.

The Bernoulli mixture model that we consider is given by

$$p(\mathbf{z}|\boldsymbol{\alpha}) = \text{Multinomial}_{K,1}(\mathbf{z}; \boldsymbol{\alpha}), \quad (15.86)$$

$$p(\mathbf{x}|\mathbf{z}, \{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{k=1}^K \{\text{Bern}_M(\mathbf{x}; \boldsymbol{\mu}_k)\}^{z_k}, \quad (15.87)$$

$$p(\boldsymbol{\alpha}|\boldsymbol{\phi}) = \text{Dirichlet}_K(\boldsymbol{\alpha}; (\boldsymbol{\phi}, \dots, \boldsymbol{\phi})^\top), \quad (15.88)$$

$$p(\boldsymbol{\mu}_k|\xi) = \prod_{m=1}^M \text{Beta}(\mu_{km}; \xi, \xi), \quad (15.89)$$

where $\boldsymbol{\phi} > 0$ and $\xi > 0$ are hyperparameters. Under the constraint, $r(\mathcal{H}, \mathbf{w}) = r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w})$, the VB posteriors are given as follows:

$$r(\{\mathbf{z}^{(n)}\}_{n=1}^N, \boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K) = r_z(\{\mathbf{z}^{(n)}\}_{n=1}^N)r_\alpha(\boldsymbol{\alpha})r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K),$$

$$r_z(\{\mathbf{z}^{(n)}\}_{n=1}^N) = \prod_{n=1}^N \text{Multinomial}_{K,1}\left(\mathbf{z}^{(n)}; \widehat{\mathbf{z}}^{(n)}\right),$$

$$r_\alpha(\boldsymbol{\alpha}) = \text{Dirichlet}\left(\boldsymbol{\alpha}; (\widehat{\boldsymbol{\phi}}_1, \dots, \widehat{\boldsymbol{\phi}}_K)^\top\right),$$

$$r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{k=1}^K \prod_{m=1}^M \text{Beta}\left(\mu_{km}; \widehat{a}_{km}, \widehat{b}_{km}\right).$$

The variational parameters $\{z^{(n)}\}_{n=1}^N, \{\widehat{\phi}_k\}_{k=1}^K, \{\{\widehat{a}_{km}\}_{m=1}^M, \{\widehat{b}_{km}\}_{m=1}^M\}_{k=1}^K$ minimize the free energy,

$$\begin{aligned} F = & \sum_{n=1}^N \sum_{k=1}^K \widehat{z}_k^{(n)} \log \widehat{z}_k^{(n)} + \log \left(\frac{\Gamma(\sum_{k=1}^K \widehat{\phi}_k)}{\prod_{k=1}^K \Gamma(\widehat{\phi}_k)} \right) - \log \left(\frac{\Gamma(K\phi)}{(\Gamma(\phi))^K} \right) \\ & + \sum_{k=1}^K (\widehat{\phi}_k - \phi - \bar{N}_k) (\Psi(\widehat{\phi}_k) - \Psi(\sum_{k'=1}^K \widehat{\phi}_{k'})) \\ & + \sum_{k=1}^K \sum_{m=1}^M \left\{ \log \left(\frac{\Gamma(\widehat{a}_{km} + \widehat{b}_{km})}{\Gamma(\widehat{a}_{km})\Gamma(\widehat{b}_{km})} \right) - \log \left(\frac{\Gamma(2\xi)}{(\Gamma(\xi))^2} \right) \right. \\ & + (\widehat{a}_{km} - \xi - \bar{N}_k \bar{x}_{km}) (\Psi(\widehat{a}_{km}) - \Psi(\widehat{a}_{km} + \widehat{b}_{km})) \\ & \left. + (\widehat{b}_{km} - \xi - \bar{N}_k (1 - \bar{x}_{km})) (\Psi(\widehat{b}_{km}) - \Psi(\widehat{a}_{km} + \widehat{b}_{km})) \right\}, \end{aligned}$$

where

$$\begin{aligned} \bar{N}_k &= \sum_{n=1}^N \langle z_k^{(n)} \rangle_{r_{\mathcal{H}}(\mathcal{H})}, \\ \bar{x}_{km} &= \frac{1}{\bar{N}_k} \sum_{n=1}^N \langle z_k^{(n)} \rangle_{r_{\mathcal{H}}(\mathcal{H})} x_m^{(n)}, \end{aligned}$$

for $k = 1, \dots, K$ and $m = 1, \dots, M$. The stationary condition of the free energy yields

$$\widehat{z}_k^{(n)} = \frac{\bar{z}_k^{(n)}}{\sum_{k'=1}^K \bar{z}_{k'}^{(n)}}, \quad (15.90)$$

$$\widehat{\phi}_k = \bar{N}_k + \phi, \quad (15.91)$$

$$\widehat{a}_{km} = \bar{N}_k \bar{x}_{km} + \xi, \quad (15.92)$$

$$\widehat{b}_{km} = \bar{N}_k (1 - \bar{x}_{km}) + \xi, \quad (15.93)$$

where

$$\begin{aligned} \bar{z}_k^{(n)} &= \exp \left(\Psi(\widehat{\phi}_k) - \Psi(\sum_{k'=1}^K \widehat{\phi}_{k'}) + \sum_{m=1}^M \left\{ x_m^{(n)} (\Psi(\widehat{a}_{km}) - \Psi(\widehat{a}_{km} + \widehat{b}_{km})) \right. \right. \\ & \left. \left. + (1 - x_m^{(n)}) (\Psi(\widehat{b}_{km}) - \Psi(\widehat{a}_{km} + \widehat{b}_{km})) \right\} \right). \end{aligned} \quad (15.94)$$

We assume the following condition.

Assumption 15.5 For $0 \leq K_1^* \leq K_0^* \leq K$, the true distribution $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{w}^*)$ is represented by K_0^* components and the parameter is given by $\mathbf{w}^* = \{\alpha_k^*, \boldsymbol{\mu}_k^*\}_{k=1}^{K_0^*}$.

$$q(\mathbf{x}) = p(\mathbf{x}|\mathbf{w}^*) = \sum_{k=1}^{K_0^*} \alpha_k^* \text{Bern}_M(\mathbf{x}; \boldsymbol{\mu}_k^*),$$

where

$$0 < \mu_{km}^* < 1 \quad (1 \leq k \leq K_1^*),$$

$$\mu_{km}^* = 0 \text{ or } 1 \quad (K_1^* + 1 \leq k \leq K_0^*).$$

We define $\Delta K^* = K_0^* - K_1^*$.

Let \widehat{K}_0 be the number of components satisfying $\bar{N}_k/N = \Omega_p(1)$ and \widehat{K}_1 be the number of components satisfying $\bar{x}_{km} = \Omega_p(1)$ and $1 - \bar{x}_{km} = \Omega_p(1)$ for all $m = 1, \dots, M$. Then, for $\Delta\widehat{K} \equiv \widehat{K}_0 - \widehat{K}_1$ components, it holds that $\bar{N}_k/N = \Omega_p(1)$ and $\bar{x}_{km} = o_p(1)$ or $1 - \bar{x}_{km} = o_p(1)$. Hence, the \widehat{K}_1 components with $\bar{N}_k/N = \Omega_p(1)$ and $\bar{x}_{km} = \Omega_p(1)$ are said to be “nondeterministic” and the $\Delta\widehat{K}$ components are said to be “deterministic,” respectively. We have the following theorem.

Theorem 15.15 *The relative free energy of the Bernoulli mixture model satisfies*

$$\begin{aligned} \widetilde{F}^{\text{VB}}(\mathcal{D}) &= F^{\text{VB}}(\mathcal{D}) - NS_N(\mathcal{D}) \\ &= \left\{ \left(\frac{M+1}{2} - \phi \right) \widehat{K}_1 + \left(\frac{1}{2} - \phi + M\xi \right) \Delta\widehat{K} + K\phi - \frac{1}{2} \right\} \log N + \Omega_p(N\widehat{J}) + O_p(1), \end{aligned}$$

where $\widehat{J} = 1$ if $\widehat{K}_1 < K_1^*$ or $\Delta\widehat{K} < \Delta K^*$ and otherwise $\widehat{J} = 0$.

The proof of Theorem 15.15 is shown after the next theorem. The following theorem claims that the numbers of deterministic and nondeterministic components are essentially determined by the hyperparameters.

Theorem 15.16 *The estimated numbers of components \widehat{K}_0 and \widehat{K}_1 of the Bernoulli mixture model are determined as follows:*

- (1) If $\frac{M+1}{2} - \phi > 0$ and $\frac{1}{2} - \phi + M\xi > 0$, then $\widehat{K}_1 = K_1^*$ and $\Delta\widehat{K} = \Delta K^*$.
- (2) If $\frac{M+1}{2} - \phi > 0$ and $\frac{1}{2} - \phi + M\xi < 0$, then $\widehat{K}_1 = K_1^*$ and $\Delta\widehat{K} = K - K_1^*$.
- (3) If $\frac{M+1}{2} - \phi < 0$ and $\frac{1}{2} - \phi + M\xi > 0$, then $\widehat{K}_1 = K - \Delta K^*$ and $\Delta\widehat{K} = \Delta K^*$.
- (4) If $\frac{M+1}{2} - \phi < 0$ and $\frac{1}{2} - \phi + M\xi < 0$, and
 - (a) if $\xi > \frac{1}{2}$, then $\widehat{K}_1 = K - \Delta K^*$ and $\Delta\widehat{K} = \Delta K^*$.
 - (b) if $\xi < \frac{1}{2}$, then $\widehat{K}_1 = K_1^*$ and $\Delta\widehat{K} = K - K_1^*$.

Proof Minimizing the coefficient of the relative free energy with respect to \widehat{K}_1 and $\Delta\widehat{K}$ under the constraint that the true distribution is realizable, i.e., $\widehat{K}_1 \geq K_1^*$ and $\Delta\widehat{K} \geq \Delta K^*$, we obtain the theorem. \square

Proof of Theorem 15.15

From Lemma 15.1, we first evaluate $R = \text{KL}(r_w(\mathbf{w})||p(\mathbf{w}))$. The inequalities of the di-gamma and log-gamma functions in Eqs. (15.8) and (15.9) yield that

$$\begin{aligned} R &= \sum_{k=1}^K \left(\frac{1}{2} - \phi \right) \log(\bar{N}_k + \phi) + \left(K\phi - \frac{1}{2} \right) \log(N + K\phi) \\ &\quad + \sum_{k=1}^K \sum_{m=1}^M \left\{ \left(\frac{1}{2} - \xi \right) \log(\bar{N}_k \bar{x}_{km} + \xi) + \left(\frac{1}{2} - \xi \right) \log(\bar{N}_k (1 - \bar{x}_{km}) + \xi) \right\} \\ &\quad + \sum_{k=1}^K M \left(2\xi - \frac{1}{2} \right) \log(\bar{N}_k + 2\xi) + O_p(1). \end{aligned}$$

We consider variational parameters in which \widehat{K}_0 components are active, i.e., $\bar{N}_k = \Omega_p(N)$. Furthermore, without loss of generality, we can assume that $0 < \bar{x}_{km} < 1$ ($1 \leq m \leq M$) for \widehat{K}_1 nondeterministic components and $\bar{x}_{km} = O_p(1/N)$ ($1 \leq m \leq M$) for $\Delta\widehat{K}$ deterministic components. Putting such variational parameters into the preceding expression, we have the asymptotic form in the theorem.

Lemmas 15.3 and 15.4 imply that if $\widehat{K}_1 < K_1^*$ or $\Delta\widehat{K} < \Delta K^*$, $\widetilde{Q} = -\log C_{\mathcal{H}} - NS_N(\mathcal{D}) = \Omega_p(N)$, and otherwise $\widetilde{Q} = O_p(1)$. Thus, we obtain the theorem.

16

Asymptotic VB Theory of Other Latent Variable Models

In this chapter, we proceed to asymptotic analyses of VB learning in other latent variable models discussed in Section 4.2, namely, Bayesian networks, hidden Markov models, probabilistic context-free grammar, and latent Dirichlet allocation.

16.1 Bayesian Networks

In this section, we analyze the VB free energy of the following *Bayesian network* model (Watanabe et al., 2009), introduced in Section 4.2.1:

$$\begin{aligned}
 p(\mathbf{x}|\mathbf{w}) &= \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x}, \mathbf{z}|\mathbf{w}), \\
 p(\mathbf{x}, \mathbf{z}|\mathbf{w}) &= p(\mathbf{x}|\mathbf{b}_z) \prod_{k=1}^K \prod_{i=1}^{T_k} a_{(k,i)}^{z_{k,i}}, \\
 p(\mathbf{x}|\mathbf{b}_z) &= \prod_{j=1}^M \prod_{l=1}^{Y_j} b_{(j,l|z)}^{x_{j,l}}, \\
 p(\mathbf{w}) &= \left\{ \prod_{k=1}^K p(\mathbf{a}_k|\phi) \right\} \left\{ \prod_{\mathbf{z} \in \mathcal{Z}} \prod_{j=1}^M p(\mathbf{b}_{j|z}|\xi) \right\}, \\
 p(\mathbf{a}_k|\phi) &= \text{Dirichlet}_{T_k}(\mathbf{a}_k; (\phi, \dots, \phi)^\top), \\
 p(\mathbf{b}_{j|z}|\xi) &= \text{Dirichlet}_{Y_j}(\mathbf{b}_{j|z}; (\xi, \dots, \xi)^\top),
 \end{aligned} \tag{16.1}$$

where $\phi > 0$ and $\xi > 0$ are hyperparameters. Here, $\mathcal{Z} = \{(z_1, \dots, z_K); z_k \in \{\mathbf{e}_i\}_{i=1}^{T_k}, k = 1, \dots, K\}$, and $z_k \in \{\mathbf{e}_i\}_{i=1}^{T_k}$ is the *one-of-K representation*, i.e., $z_{k,i} = 1$ for some $i \in \{1, \dots, T_k\}$ and $z_{k,j} = 0$ for $j \neq i$. Also, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_M)$ for $\mathbf{x}_j \in \{\mathbf{e}_l\}_{l=1}^{Y_j}$. The number of the parameters of this model is

$$D = M_{\text{obs}} \prod_{k=1}^K T_k + \sum_{k=1}^K (T_k - 1), \quad (16.2)$$

where

$$M_{\text{obs}} = \sum_{j=1}^M (Y_j - 1).$$

Under the constraint, $r(\mathcal{H}, \mathbf{w}) = r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w})$, the VB posteriors are given by

$$r_w(\mathbf{w}) = \left\{ \prod_{k=1}^K r_a(\mathbf{a}_k) \right\} \left\{ \prod_{z \in \mathcal{Z}} \prod_{j=1}^M r_b(\mathbf{b}_{j|z}) \right\},$$

$$r_a(\mathbf{a}_k) = \text{Dirichlet}_{T_k}(\mathbf{a}_k; \widehat{\boldsymbol{\phi}}_k), \quad (16.3)$$

$$r_b(\mathbf{b}_{j|z}) = \text{Dirichlet}_{Y_j}(\mathbf{b}_{j|z}; \widehat{\boldsymbol{\xi}}_{j|z}), \quad (16.4)$$

$$r_{\mathcal{H}}(\mathcal{H}) = \prod_{n=1}^N r_z(z^{(n)}),$$

where

$$\begin{aligned} r_z(z^{(n)} = z) &\propto \exp \left(\sum_{k=1}^K \left\{ \Psi(\widehat{\boldsymbol{\phi}}_{(k,i_k)}) - \Psi \left(\sum_{i'_k=1}^{T_k} \widehat{\boldsymbol{\phi}}_{(k,i'_k)} \right) \right\} \right. \\ &\quad \left. + \sum_{j=1}^M \left\{ \Psi(\widehat{\boldsymbol{\xi}}_{(j,l_j^{(n)}|z)}) - \Psi \left(\sum_{l'=1}^{Y_j} \widehat{\boldsymbol{\xi}}_{(j,l'|z)} \right) \right\} \right) \end{aligned} \quad (16.5)$$

for $\mathbf{z} = (\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_K})$ and $\mathbf{x}_j^{(n)} = \mathbf{e}_{l_j^{(n)}}$. The free energy is given by

$$\begin{aligned} F &= \sum_{k=1}^K \left\{ \log \left(\frac{\Gamma(\sum_{i=1}^{T_k} \widehat{\boldsymbol{\phi}}_{(k,i)})}{\prod_{i=1}^{T_k} \Gamma(\widehat{\boldsymbol{\phi}}_{(k,i)})} \right) - \log \left(\frac{\Gamma(T_k \phi)}{(\Gamma(\phi))^{T_k}} \right) \right. \\ &\quad \left. + \sum_{i=1}^{T_k} (\widehat{\boldsymbol{\phi}}_{(k,i)} - \phi - \overline{N}_{(k,i)}^z) (\Psi(\widehat{\boldsymbol{\phi}}_{(k,i)}) - \Psi(\sum_{i'=1}^{T_k} \widehat{\boldsymbol{\phi}}_{(k,i')})) \right\} \\ &\quad + \sum_{z \in \mathcal{Z}} \sum_{j=1}^M \left\{ \log \left(\frac{\Gamma(\sum_{l=1}^{Y_j} \widehat{\boldsymbol{\xi}}_{(j,l|z)})}{\prod_{l=1}^{Y_j} \Gamma(\widehat{\boldsymbol{\xi}}_{(j,l|z)})} \right) - \log \left(\frac{\Gamma(Y_j \xi)}{(\Gamma(\xi))^{Y_j}} \right) \right. \\ &\quad \left. + \sum_{l=1}^{Y_j} (\widehat{\boldsymbol{\xi}}_{(j,l|z)} - \xi - \overline{N}_{(j,l|z)}^x) (\Psi(\widehat{\boldsymbol{\xi}}_{(j,l|z)}) - \Psi(\sum_{l'=1}^{Y_j} \widehat{\boldsymbol{\xi}}_{(j,l'|z)})) \right\} \\ &\quad + \sum_{n=1}^N \sum_{z \in \mathcal{Z}} r_z(z^{(n)} = z) \log r_z(z^{(n)} = z), \end{aligned}$$

where

$$\begin{aligned}\overline{N}_{(k,i_k)}^z &= \sum_{n=1}^N \left\langle z_{k,i_k}^{(n)} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})}, \\ \overline{N}_{(j,l_j|z)}^x &= \sum_{n=1}^N x_{j,l_j}^{(n)} r_z(z^{(n)} = z),\end{aligned}$$

for

$$r_z(z^{(n)} = (\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_K})) = \left\langle \prod_{k=1}^K z_{k,i_k}^{(n)} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})}. \quad (16.6)$$

We assume the following condition:

Assumption 16.1 *The true distribution $q(\mathbf{x})$ can be expressed by a Bayesian network with H hidden nodes, each of which has S_k states, for $H \leq K$, i.e.,*

$$q(\mathbf{x}) = p(\mathbf{x}|\mathbf{w}^*) = \sum_{\mathbf{z} \in \mathcal{Z}^*} p(\mathbf{x}, \mathbf{z}|\mathbf{w}^*) = \sum_{\mathbf{z} \in \mathcal{Z}^*} p(\mathbf{x}|\mathbf{b}_z^*) \prod_{k=1}^H \prod_{i=1}^{S_k} \{a_{(k,i)}^*\}^{z_{k,i}},$$

where

$$p(\mathbf{x}|\mathbf{b}_z^*) = \prod_{j=1}^M \prod_{l=1}^{Y_j} \{b_{(j,l|z)}^*\}^{x_{j,l}}$$

for $\mathbf{z} \in \mathcal{Z}^* = \{(z_1, \dots, z_H); z_k \in \{\mathbf{e}_i\}_{i=1}^{S_k}, k = 1, \dots, H\}$. The true parameters $\mathbf{w}^* = \{\{\mathbf{a}_k^*\}_{k=1}^H, \{\mathbf{b}_z^*\}_{z \in \mathcal{Z}^*}\}$ are given by

$$\begin{aligned}\mathbf{a}_k^* &= \{a_{(k,i)}^*; 1 \leq i \leq S_k\} \quad (k = 1, \dots, H), \\ \mathbf{b}_z^* &= \{b_{(j,l|z)}^*\}_{j=1}^M \quad (z \in \mathcal{Z}^*), \\ \mathbf{b}_{j|z}^* &= \{b_{(j,l|z)}^*; 1 \leq l \leq Y_j\} \quad (j = 1, \dots, M).\end{aligned}$$

For $k > H$, we define $S_k = 1$.

The true distribution can be realized by the model, i.e., the model is given by Eq. (16.1), where $T_k \geq S_k$ holds for $k = 1, \dots, H$. We assume that the true distribution is the smallest in the sense that it cannot be realized by any model with a smaller number of hidden units and with a smaller number of the states of each hidden unit.

Under this condition, we prove the following theorem, which evaluates the relative VB free energy. The proof will appear in the next section.

Theorem 16.1 *The relative VB free energy of the Bayesian network model satisfies*

$$\tilde{F}^{\text{VB}}(\mathcal{D}) = F^{\text{VB}}(\mathcal{D}) - NS_N(\mathcal{D}) = \lambda'_{\text{BN}}^{\text{VB}} \log N + O_p(1),$$

where

$$\lambda'_{\text{BN}}^{\text{VB}} = \phi \sum_{k=1}^K T_k - \frac{K}{2} + \min_{\{u_k\}} \left\{ \frac{M_{\text{obs}}}{2} \prod_{k=1}^K u_k - \left(\phi - \frac{1}{2} \right) \sum_{k=1}^K u_k \right\}. \quad (16.7)$$

The minimum is taken over the set of positive integers $\{u_k; S_k \leq u_k \leq T_k\}_{k=1}^K$.

If $K = 1$, this is reduced to the case of the naive Bayesian networks whose Bayes free energy or stochastic complexity has been evaluated (Yamazaki and Watanabe, 2003a; Rusakov and Geiger, 2005). Bounds for their VB free energy have also been obtained (Watanabe and Watanabe, 2004, 2005, 2006).

The coefficient $\lambda'_{\text{BN}}^{\text{VB}}$ is given by the solution of the minimization problem in Eq. (16.7). We present a few exemplary cases as corollaries in this section.

By taking $u_k = S_k$ for $1 \leq k \leq H$ and $u_k = 1$ for $H+1 \leq k \leq K$, we obtain the following upper-bound for the VB free energy (Watanabe et al., 2006). This bound is tight if $\phi \leq (1 + M_{\text{obs}} \min_{1 \leq k \leq K} \{S_k\})/2$.

Corollary 16.2 *It holds that*

$$\tilde{F}^{\text{VB}}(\mathcal{D}) \leq \lambda'_{\text{BN}}^{\text{VB}} \log N + O_p(1), \quad (16.8)$$

where

$$\lambda'_{\text{BN}}^{\text{VB}} = \phi \sum_{k=1}^K T_k - \phi K + \left(\phi - \frac{1}{2} \right) H + \left(\frac{1}{2} - \phi \right) \sum_{k=1}^H S_k + \frac{M_{\text{obs}}}{2} \prod_{k=1}^H S_k. \quad (16.9)$$

If $K = H = 2$, that is, the true network and the model both have two hidden nodes, solving the minimization problem in Eq. (16.7) gives the following corollary. Suppose $S_1 \geq S_2$ and $T_1 \geq T_2$.

Corollary 16.3 *If $K = H = 2$,*

$$\tilde{F}^{\text{VB}}(\mathcal{D}) = \lambda'_{\text{BN}}^{\text{VB}} \log N + O_p(1), \quad (16.10)$$

where

$$\begin{aligned} \lambda'_{\text{BN}}^{\text{VB}} &= \begin{cases} (T_1 - S_1 + T_2 - S_2)\phi + \frac{M_{\text{obs}}}{2} S_1 S_2 + \frac{S_1 + S_2}{2} - 1 & (0 < \phi \leq \frac{1+S_2 M_{\text{obs}}}{2}), \\ (T_2 - S_2)\phi + \frac{M_{\text{obs}}}{2} T_1 S_2 + \frac{T_1 + S_2}{2} - 1 & (\frac{1+S_2 M_{\text{obs}}}{2} < \phi \leq \frac{1+T_1 M_{\text{obs}}}{2}), \\ \frac{M_{\text{obs}}}{2} T_1 T_2 + \frac{T_1 + T_2}{2} - 1 & (\frac{1+T_1 M_{\text{obs}}}{2} < \phi). \end{cases} \end{aligned} \quad (16.11)$$

The leading term of the relative Bayes free energy of regular statistical models is given by $(D/2) \log N$ (Schwarz, 1978), where D is the number of parameters in Eq. (16.2). Corollary 16.3 claims that the coefficient $\lambda_{\text{BN}}^{\text{VB}}$ of the leading term is smaller than $D/2$ when $\phi \leq \frac{1+T_1 M_{\text{obs}}}{2}$.

Proof of Theorem 16.1

From Lemma 15.1, we can rewrite the free energy as follows:

$$F^{\text{VB}}(\mathcal{D}) = \min_{r_w(\mathbf{w})} [R + Q], \quad (16.12)$$

where

$$\begin{aligned} R &= \text{KL}(r_w(\mathbf{w}) || p(\mathbf{w})), \\ Q &= -\log C_{\mathcal{H}} = -\log \sum_{\mathcal{H}} \langle \log p(\mathcal{D}, \mathcal{H} | \mathbf{w}) \rangle_{r_w(\mathbf{w})}. \end{aligned}$$

From Eqs. (16.3), (16.4), and (16.5), we obtain Q and R in Eq. (16.12) as follows:

$$\begin{aligned} Q &= -\sum_{n=1}^N \log \sum_{z^{(n)}} \langle \log p(\mathbf{x}^{(n)}, z^{(n)} | \mathbf{w}) \rangle_{r_w(\mathbf{w})} \\ &= -\sum_{n=1}^N \log \left(\sum_{\mathbf{z}=(e_{i_1}, \dots, e_{i_K})} \exp \left(\sum_{k=1}^K \left\{ \Psi(\bar{N}_{(k,i_k)}^z + \phi) - \Psi(N + T_k \phi) \right\} \right. \right. \\ &\quad \left. \left. + \sum_{j=1}^M \sum_{l=1}^{Y_j} x_{j,l}^{(n)} \left\{ \Psi(\bar{N}_{(j,l|z)}^x + \xi) - \Psi(\bar{N}_z^x + Y_j \xi) \right\} \right) \right), \end{aligned} \quad (16.13)$$

and

$$\begin{aligned} R &= \sum_{k=1}^K \text{KL}(r_a(\mathbf{a}_k) || p(\mathbf{a}_k | \phi)) + \sum_z \sum_{j=1}^M \text{KL}(r_b(\mathbf{b}_{j|z}) || p(\mathbf{b}_{j|z} | \xi)) \\ &= \sum_{k=1}^K \left[\sum_{i=1}^{T_k} \left\{ \bar{N}_{(k,i)}^z \Psi(\bar{N}_{(k,i)}^z + \phi) - \log \Gamma(\bar{N}_{(k,i)}^z + \phi) \right\} \right. \\ &\quad \left. - N \Psi(N + T_k \phi) + \log \Gamma(N + T_k \phi) + \log \frac{\Gamma(\phi)^{T_k}}{\Gamma(T_k \phi)} \right] \\ &\quad + \sum_z \sum_{j=1}^M \left[\sum_{l=1}^{Y_j} \left\{ \bar{N}_{(j,l|z)}^x \Psi(\bar{N}_{(j,l|z)}^x + \xi) - \log \Gamma(\bar{N}_{(j,l|z)}^x + \xi) \right\} \right. \\ &\quad \left. - \bar{N}_z^x \Psi(\bar{N}_z^x + Y_j \xi) + \log \Gamma(\bar{N}_z^x + Y_j \xi) + \log \frac{\Gamma(\xi)^{Y_j}}{\Gamma(Y_j \xi)} \right]. \end{aligned} \quad (16.14)$$

Furthermore, by using the inequalities for the di-gamma and log-gamma functions in Eqs. (15.8) and (15.9), we can bound Q as follows:

$$Q \leq - \sum_{n=1}^N \log \left(\sum_{\mathbf{z}=(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_K})} \exp \left(\sum_{k=1}^K \left\{ \log \frac{\bar{N}_{(k,i_k)}^z + \phi}{N + T_k \phi} - \frac{1}{\bar{N}_{(k,i_k)}^z + \phi} + \frac{1}{2(N + T_k \phi)} \right\} \right. \right. \\ \left. \left. + \sum_{j=1}^M \sum_{l=1}^{Y_j} x_{j,l}^{(n)} \left\{ \log \frac{\bar{N}_{(j,l|\mathbf{z})}^x + \xi}{\bar{N}_{\mathbf{z}}^x + Y_j \xi} - \frac{1}{\bar{N}_{(j,l|\mathbf{z})}^x + \xi} + \frac{1}{2(\bar{N}_{\mathbf{z}}^x + Y_j \xi)} \right\} \right) \right). \quad (16.15)$$

We can also evaluate R in Eq. (16.12) as follows:

$$R = \sum_{k=1}^K \left\{ \left(T_k \phi - \frac{1}{2} \right) \log (N + T_k \phi) \right\} - \sum_{k=1}^K \sum_{i=1}^{T_k} \left\{ \left(\phi - \frac{1}{2} \right) \log (\bar{N}_{(k,i)}^z + \phi) \right\} \\ + \sum_{\mathbf{z}} \sum_{j=1}^M \left\{ \left(Y_j \xi - \frac{1}{2} \right) \log (\bar{N}_{\mathbf{z}}^x + Y_j \xi) - \sum_{l=1}^{Y_j} \left(\xi - \frac{1}{2} \right) \log (\bar{N}_{(j,l|\mathbf{z})}^x + \xi) \right\} \\ + O_p(1). \quad (16.16)$$

Since $F^{\text{VB}}(\mathcal{D})$ is given as the minimum value of the function of $\{\bar{N}_{(j,l|\mathbf{z})}^x\}$, we can obtain from Eq. (16.12) an upper-bound for $F^{\text{VB}}(\mathcal{D})$ by substituting each $\bar{N}_{(j,l|\mathbf{z})}^x$ by any specific value. Therefore, let u_k be a natural number such that $S_k \leq u_k \leq T_k$ for $k = 1, \dots, K$ and consider the following $\bar{N}_{(j,l|\mathbf{z})}^x$ for each j and l :

$$\bar{N}_{(j,l|\bar{\mathbf{z}})}^x = N b_{(j,l|\bar{\mathbf{z}})}^* \prod_{k=1}^K \bar{a}_{(k,i_k)}, \quad (16.17)$$

where $\bar{\mathbf{z}} = (\mathbf{e}_{\min\{l_1, S_1\}}, \mathbf{e}_{\min\{l_2, S_2\}}, \dots, \mathbf{e}_{\min\{l_H, S_H\}})$ and

$$\bar{a}_{(k,i_k)} = \begin{cases} a_{(k,i_k)}^* & (1 \leq i_k \leq S_k - 1), \\ a_{(k,S_k)}^*/(u_k - S_k + 1) & (S_k \leq i_k \leq u_k), \\ 0 & (\text{otherwise}). \end{cases} \quad (16.18)$$

This corresponds to the case where u_k ($\geq S_k$) states of the k th hidden node are active for $k = 1, \dots, H$. Then we have $\bar{N}_{\mathbf{z}}^x = N \prod_{k=1}^K \bar{a}_{(k,i_k)}$ and $\bar{N}_{(k,i)}^z = N \bar{a}_{(k,i)}$. Substituting them into Eq. (16.16) yields

$$R = \left\{ \phi \sum_{k=1}^K T_k - \frac{K}{2} + \frac{M_{\text{obs}}}{2} \prod_{k=1}^K u_k - \left(\phi - \frac{1}{2} \right) \sum_{k=1}^K u_k \right\} \log N + O_p(1). \quad (16.19)$$

From Eq. (16.15), we obtain

$$\begin{aligned} Q &\leq - \sum_{n=1}^N \log \left(p(\mathbf{x}^{(n)} | \mathbf{w}^*) \exp \left(O_p \left(\frac{1}{N} \right) \right) \right) \\ &= NS_N(\mathcal{D}) + O_p(1). \end{aligned} \quad (16.20)$$

From Eqs. (16.12), (16.19), and (16.20), we have proved that $\tilde{F}^{\text{VB}}(\mathcal{D})$ is upper-bounded by the right-hand side of Eq. (16.19) for any data set \mathcal{D} and $\{u_k; S_k \leq u_k \leq T_k\}$.

If the number of states such that $\bar{N}_{(k,i)} = \Theta_p(N)$ is less than S_k , i.e., $u_k < S_k$ for some k , the predictive distribution $\langle p(\mathbf{x}|\mathbf{w}) \rangle_{r_w(\mathbf{w})}$ cannot approach the true distribution $p(\mathbf{x}|\mathbf{w}^*)$. Then Lemma 15.4 implies that $Q - NS_N(\mathcal{D}) = \Omega_p(N)$. Hence, minimizing the coefficient of the leading term in Eq. (16.19) under the constraints $S_k \leq u_k \leq T_k$ for all k , we complete the proof.

16.2 Hidden Markov Models

Next we analyze the VB free energy of hidden Markov models (HMMs) (Hosino et al., 2005, 2006b), introduced in Section 4.2.2. Suppose that we observe N sequences, $\mathcal{D} = \{\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}\}$, where each sequence $\mathbf{X}^{(n)} = (\mathbf{x}^{(n,1)}, \dots, \mathbf{x}^{(n,T)})$ has length T . We consider the asymptotic analysis for the VB free energy as the number of i.i.d. sample sequences tends to infinity, i.e., $N \rightarrow \infty$ while T is a fixed constant.

The model for observed and hidden sequences $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)})$ and $\mathbf{Z} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(T)})$ is given by

$$\begin{aligned} p(\mathbf{X}|\mathbf{w}) &= \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\mathbf{w}), \\ p(\mathbf{X}, \mathbf{Z}|\mathbf{w}) &= \prod_{m=1}^M b_{1,m}^{x_m^{(1)}} \prod_{t=2}^T \prod_{k=1}^K \prod_{l=1}^K a_{k,l}^{z_l^{(t)} z_k^{(t-1)}} \prod_{m=1}^M b_{k,m}^{z_k^{(t)} x_m^{(t)}}, \\ p(\mathbf{A}|\phi) &= \prod_{k=1}^K \text{Dirichlet}_K \left(\bar{\mathbf{a}}_k; (\phi, \dots, \phi)^\top \right), \\ p(\mathbf{B}|\xi) &= \prod_{k=1}^K \text{Dirichlet}_M \left(\bar{\mathbf{b}}_k; (\xi, \dots, \xi)^\top \right), \end{aligned} \quad (16.21)$$

where $\phi > 0$ and $\xi > 0$ are hyperparameters. Let $\mathcal{H} = \{\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(N)}\}$ be the set of hidden sequences. Under the constraint, $r(\mathcal{H}, \mathbf{w}) = r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w})$, the VB posteriors are given by

$$\begin{aligned}
r_w(\mathbf{w}) &= r_A(\mathbf{A})r_B(\mathbf{B}), \\
r_A(\mathbf{A}) &= \prod_{k=1}^K \text{Dirichlet}_K \left(\tilde{\mathbf{a}}_k; (\widehat{\phi}_{k,1}, \dots, \widehat{\phi}_{k,K})^\top \right), \\
r_B(\mathbf{B}) &= \prod_{k=1}^K \text{Dirichlet}_M \left(\tilde{\mathbf{b}}_k; (\widehat{\xi}_{k,1}, \dots, \widehat{\xi}_{k,M})^\top \right), \\
r_{\mathcal{H}}(\mathcal{H}) &= \prod_{n=1}^N r_Z(\mathbf{Z}^{(n)}), \\
r_Z(\mathbf{Z}^{(n)}) &= \frac{1}{C_{\mathbf{Z}^{(n)}}} \exp \left(\sum_{t=2}^T \sum_{k=1}^K \sum_{l=1}^K z_k^{(n,t)} z_l^{(n,t-1)} \left\{ \Psi(\widehat{\phi}_{k,l}) - \Psi \left(\sum_{l'=1}^K \widehat{\phi}_{k,l'} \right) \right\} \right. \\
&\quad \left. + \sum_{t=1}^T \sum_{k=1}^K \sum_{m=1}^M z_k^{(n,t)} x_m^{(t)} \left\{ \Psi(\widehat{\xi}_{k,m}) - \Psi \left(\sum_{m'=1}^M \widehat{\xi}_{k,m'} \right) \right\} \right),
\end{aligned}$$

where $C_{\mathbf{Z}^{(n)}}$ is the normalizing constant. After the substitution of Eq. (15.5), the free energy is given by

$$\begin{aligned}
F &= \sum_{k=1}^K \left\{ \log \left(\frac{\Gamma(\sum_{l=1}^K \widehat{\phi}_{k,l})}{\prod_{l=1}^K \Gamma(\widehat{\phi}_{k,l})} \right) + \sum_{l=1}^K (\widehat{\phi}_{k,l} - \phi) (\Psi(\widehat{\phi}_{k,l}) - \Psi(\sum_{l'=1}^K \widehat{\phi}_{k,l'})) \right. \\
&\quad \left. + \log \left(\frac{\Gamma(\sum_{m=1}^M \widehat{\xi}_{k,m})}{\prod_{m=1}^M \Gamma(\widehat{\xi}_{k,m})} \right) + \sum_{m=1}^M (\widehat{\xi}_{k,m} - \xi) (\Psi(\widehat{\xi}_{k,m}) - \Psi(\sum_{m'=1}^M \widehat{\xi}_{k,m'})) \right\} \\
&\quad - K \log \left(\frac{\Gamma(K\phi)}{(\Gamma(\phi))^K} \right) - K \log \left(\frac{\Gamma(M\xi)}{(\Gamma(\xi))^M} \right) - \sum_{n=1}^N \log C_{\mathbf{Z}^{(n)}}.
\end{aligned}$$

The variational parameters satisfy

$$\begin{aligned}
\widehat{\phi}_{k,l} &= \overline{N}_{k,l}^{[z]} + \phi, \\
\widehat{\xi}_{k,m} &= \overline{N}_{k,m}^{[x]} + \xi,
\end{aligned}$$

for the expected sufficient statistics defined by

$$\begin{aligned}
\overline{N}_{k,l}^{[z]} &= \sum_{n=1}^N \sum_{t=2}^T \langle z_l^{(n,t)} z_k^{(n,t-1)} \rangle_{r_{\mathcal{H}}(\mathcal{H})}, \\
\overline{N}_{k,m}^{[x]} &= \sum_{n=1}^N \sum_{t=1}^T \langle z_k^{(n,t)} \rangle_{r_{\mathcal{H}}(\mathcal{H})} x_m^{(n,t)}.
\end{aligned}$$

We assume the following condition.

Assumption 16.2 *The true distribution $q(X)$ has K_0 hidden states and emits M -valued discrete symbols:*

$$q(X) = p(X|\boldsymbol{w}^*) = \sum_{\mathbf{z}} \prod_{m=1}^M (b_{1m}^*)^{x_m^{(1)}} \prod_{t=2}^T \prod_{k=1}^{K_0} \prod_{l=1}^{K_0} (a_{kl}^*)^{z_l^{(t)} z_k^{(t-1)}} \prod_{m=1}^M (b_{km}^*)^{z_k^{(t)} x_m^{(t)}}, \quad (16.22)$$

where $\sum_{\mathbf{z}}$ is taken over all possible values of the hidden variables. Moreover, the true parameter is defined by

$$\boldsymbol{w}^* = (\boldsymbol{A}^*, \boldsymbol{B}^*) = ((a_{kl}^*), (b_{km}^*)),$$

where $\boldsymbol{A}^* \in \mathbb{R}^{K_0 \times K_0}$ and $\boldsymbol{B}^* \in \mathbb{R}^{K_0 \times m}$. The number of hidden states K_0 of the true distribution is the smallest under this parameterization (Ito et al., 1992) and all parameters $\{a_{kl}^*, b_{km}^*\}$ are strictly positive:

$$\boldsymbol{w}^* = ((a_{kl}^* > 0), (b_{km}^* > 0)) \quad (1 \leq k, l \leq K_0, 1 \leq m \leq M).$$

The statistical model given by Eq. (16.21) can attain the true distribution, thus the model has K ($\geq K_0$) hidden states.

Under this assumption, the next theorem evaluates the relative VB free energy. Here $S_N(\mathcal{D}) = -\frac{1}{N} \sum_{n=1}^N \log p(X^{(n)}|\boldsymbol{w}^*)$ is the empirical entropy of the true distribution (16.22).

Theorem 16.4 *The relative VB free energy of HMMs satisfies*

$$\tilde{F}^{\text{VB}}(\mathcal{D}) = F^{\text{VB}}(\mathcal{D}) - NS_N(\mathcal{D}) = \lambda'_{\text{HMM}}^{\text{VB}} \log N + O_p(1),$$

where

$$\lambda'_{\text{HMM}}^{\text{VB}} = \begin{cases} \frac{K_0(K_0-1)+K_0(M-1)}{2} + K_0(K-K_0)\phi & \left(0 < \phi \leq \frac{K_0+K+M-2}{2K_0}\right), \\ \frac{K(K-1)+K(M-1)}{2} & \left(\frac{K_0+K+M-2}{2K_0} < \phi\right). \end{cases} \quad (16.23)$$

Proof As in the models discussed in the previous sections, we evaluate the KL divergence from the posterior distribution to the prior distribution of parameters:

$$\begin{aligned} R &= \text{KL}(r_w(\boldsymbol{w})||p(\boldsymbol{w})) \\ &= \sum_{k=1}^K \left[\log \Gamma(\bar{N}_k + K\phi) - \bar{N}_k \Psi(\bar{N}_k + K\phi) \right. \\ &\quad \left. - \sum_{l=1}^K \left\{ \log \Gamma(\bar{N}_{k,l}^{[z]} + \phi) - \bar{N}_{k,l}^{[z]} \Psi(\bar{N}_{k,l}^{[z]} + \phi) \right\} \right] \end{aligned}$$

$$\begin{aligned}
& + \log \Gamma(\bar{N}_k + M\xi) - \bar{N}_k \Psi(\bar{N}_k + M\xi) \\
& - \sum_{m=1}^M \left\{ \log \Gamma(\bar{N}_{k,m}^{[x]} + \xi) - \bar{N}_{k,m}^{[x]} \Psi(\bar{N}_{k,m}^{[x]} + \xi) \right\} + O_p(1). \tag{16.24}
\end{aligned}$$

Using the inequalities of the di-gamma and the log-gamma functions in Eqs. (15.8) and (15.9), we have

$$\begin{aligned}
R = & \sum_{k=1}^K \left[\left(K\phi - \frac{1}{2} \right) \log(\bar{N}_k + K\phi) - \sum_{l=1}^K \left(\phi - \frac{1}{2} \right) \log(\bar{N}_{k,l}^{[z]} + \phi) \right. \\
& \left. + \left(M\xi - \frac{1}{2} \right) \log(\bar{N}_k + M\xi) - \sum_{m=1}^M \left(\xi - \frac{1}{2} \right) \log(\bar{N}_{k,m}^{[x]} + \xi) \right] + O_p(1). \tag{16.25}
\end{aligned}$$

We divide the sum over k and l in Eq. (16.25) to the necessary K_0 and redundant $K - K_0$ terms. Moreover, we assume that additional l ($0 \leq l \leq K - K_0$) hidden states are used, i.e., having $\bar{N}_k = \Theta_p(N)$.

$$\frac{R}{\log N} = \sum_{k=1}^{K_0} \left\{ \left(K\phi + \frac{M}{2} - 1 \right) - \sum_{l=1}^{K_0} \left(\phi - \frac{1}{2} \right) \right\} + g(l) + O_p\left(\frac{1}{\log N}\right), \tag{16.26}$$

where $g(l)$ is given by

$$g(l) = \left(K\phi + \frac{M}{2} - 1 \right) l - \left(\phi - \frac{1}{2} \right) (2K_0 l + l^2).$$

If the number of states with $\bar{N}_k = \Theta_p(N)$ is less than K_0 , Lemma 15.4 implies that $\tilde{Q} = -\log C_{\mathcal{H}} - NS_N(\mathcal{D}) = \Omega_p(N)$ for data sequences in the strongly ε -typical set. Otherwise, we can upper-bound $F^{\text{VB}}(\mathcal{D}) - NS_N(\mathcal{D})$ so that $\tilde{Q} = O_p(1)$ similarly to the models in the previous sections. Hence, minimizing the right-hand side of Eq. (16.26) with respect to l , we can evaluate the VB free energy.

The minimum of $g(l)$ is achieved by

$$\begin{cases} l = 0 & \left(0 < \phi \leq \frac{K_0+K+M-2}{2K_0} \right), \\ l = K - K_0 & \left(\frac{K_0+K+M-2}{2K_0} < \phi \right). \end{cases}$$

Putting this back into Eq. (16.26), we obtain the theorem. \square

Next we consider the simple left-to-right HMMs.

Assumption 16.3 *In the simple left-to-right HMMs, transition from each hidden state is constrained to itself or the next hidden state:*

$$\{a_{k,l} = 0, l \neq \{k, k+1\}\}. \tag{16.27}$$

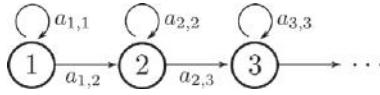


Figure 16.1 State transition diagram of a left-to-right HMM.

Thus, only $a_{k,k+1}$ is a substantial parameter in the transition probability. Figure 16.1 illustrates the state transition diagram of a left-to-right HMM.

The next theorem evaluates the relative VB free energy of the left-to-right HMM.¹

Theorem 16.5 *The relative VB free energy of the left-to-right HMM satisfies*

$$\tilde{F}^{\text{VB}}(\mathcal{D}) = \lambda'_{\text{LR-HMM}}^{\text{VB}} \log N + O_p(1),$$

where

$$\lambda'_{\text{LR-HMM}}^{\text{VB}} = \begin{cases} \frac{(K_0-1)+K_0(M-1)}{2} + \phi & (\phi \leq \frac{M(K-K_0)}{2}), \\ \frac{(K-1)+K(M-1)}{2} & (\phi > \frac{M(K-K_0)}{2}). \end{cases} \quad (16.28)$$

Proof From the constraints of the transition probabilities in Eq. (16.27), the asymptotic form of the KL divergence from the VB posterior to the prior is given by

$$\begin{aligned} R &= \text{KL}(r_w(\mathbf{w}) \| p(\mathbf{w})) \\ &= \sum_{k=1}^{K-1} \left[\left(2\phi - \frac{1}{2} \right) \log(\bar{N}_k + 2\phi) - \left(\phi - \frac{1}{2} \right) \left\{ \log(\bar{N}_{k,(k+1)}^{[z]} + \phi) + \log(\bar{N}_{k,k}^{[z]} + \phi) \right\} \right] \\ &\quad + \sum_{k=1}^K \left[\left(M\xi - \frac{1}{2} \right) \log(\bar{N}_k + M\xi) - \sum_{m=1}^M \left(\xi - \frac{1}{2} \right) \log(\bar{N}_{k,m}^{[x]} + \xi) \right] \\ &\quad + O_p(1). \end{aligned} \quad (16.29)$$

If K hidden states are used, all the variables, \bar{N}_k , $\bar{N}_{k,k}^{[z]}$, $\bar{N}_{k,(k+1)}^{[z]}$, and $\bar{N}_{k,m}^{[x]}$ are in the order of N , which leads to the asymptotic form in the theorem. If some states are not used, we assume that the $(K_0 + l)$ th state is the last state that is effectively used. More specifically, if we consider the case where \bar{N}_k , $\bar{N}_{k,k}^{[z]}$, $\bar{N}_{k,(k+1)}^{[z]}$, and $\bar{N}_{k,m}^{[x]}$ are $\Theta_p(N)$ for $K_0 + l - 1$ states and $\bar{N}_{(K_0+l),(K_0+l+1)}^{[z]} = O_p(1)$ and $\bar{N}_{(K_0+l),(K_0+l)}^{[z]} = \Theta_p(N)$ (and hence, $\bar{N}_{K_0+l} = \Theta_p(N)$), we obtain

$$\frac{R}{\log N} = \frac{K_0 - 1}{2} + \phi + K_0 \frac{M - 1}{2} + g(l) + O_p\left(\frac{1}{\log N}\right),$$

¹ This theorem is not obtained as a special case of Theorem 16.4 since some of the transition probabilities are fixed to zero and are no longer parameters.

where

$$g(l) = \frac{M}{2}l$$

for $0 \leq l \leq K - K_0$. Since the minimum of $g(l)$ is obviously obtained by $l = 0$, we obtain the theorem. \square

16.3 Probabilistic Context-Free Grammar

In this section, we asymptotically analyze the VB free energy of probabilistic context-free grammar (PCFG), introduced in Section 4.2.3, as the number N of the sequences in the training corpus $\mathcal{D} = \{X^{(1)}, \dots, X^{(N)}\}$ goes to infinity (Hosino et al., 2006a). The PCFG model is defined by

$$\begin{aligned} p(X|\mathbf{w}) &= \sum_{Z \in T(X)} p(X, Z|\mathbf{w}), \\ p(X, Z|\mathbf{w}) &= \prod_{i,j,k=1}^K (a_{i \rightarrow jk})^{c_{i \rightarrow jk}^Z} \prod_{l=1}^L \prod_{i=1}^K \prod_{m=1}^M (b_{i \rightarrow m})^{\tilde{z}_i^{(l)} x_m^0}, \\ \mathbf{w} &= \{\{\mathbf{a}_i\}_{i=1}^K, \{\mathbf{b}_i\}_{i=1}^K\}, \\ \mathbf{a}_i &= \{a_{i \rightarrow jk}\}_{j,k=1}^K \quad (1 \leq i \leq K), \\ \mathbf{b}_i &= \{b_{i \rightarrow m}\}_{m=1}^M \quad (1 \leq i \leq K), \\ p(\{\mathbf{a}_i\}_{i=1}^K | \phi) &= \prod_{i=1}^K \text{Dirichlet}_{K^2}(\mathbf{a}_i; (\phi, \dots, \phi)^\top), \\ p(\{\mathbf{b}_i\}_{i=1}^K | \xi) &= \prod_{i=1}^K \text{Dirichlet}_M(\mathbf{b}_i; (\xi, \dots, \xi)^\top), \end{aligned} \tag{16.30}$$

where $\phi > 0$ and $\xi > 0$ are hyperparameters. Here $T(X)$ is the set of derivation sequences that generate X , $c_{i \rightarrow jk}^Z$ is the count of the transition rule from the nonterminal symbol i to the pair of nonterminal symbols (j, k) appearing in the derivation sequence Z , and $\tilde{\mathbf{z}}^{(l)} = (\tilde{z}_1^{(l)}, \dots, \tilde{z}_K^{(l)})$ is the indicator of the (nonterminal) symbol generating the l th output symbol of X .

Under the constraint, $r(\mathcal{H}, \mathbf{w}) = r_{\mathcal{H}}(\mathcal{H})r_w(\mathbf{w})$, the VB posteriors are given by

$$\begin{aligned} r_w(\mathbf{w}) &= r_a(\{\mathbf{a}_i\}_{i=1}^K) r_b(\{\mathbf{b}_i\}_{i=1}^K), \\ r_a(\{\mathbf{a}_i\}_{i=1}^K) &= \prod_{i=1}^K \text{Dirichlet}_{K^2}(\mathbf{a}_i; (\widehat{\phi}_{i \rightarrow 11}, \dots, \widehat{\phi}_{i \rightarrow KK})^\top), \end{aligned}$$

$$\begin{aligned}
r_b(\{\mathbf{b}_i\}_{i=1}^K) &= \prod_{i=1}^K \text{Dirichlet}_M \left(\mathbf{b}_i; (\widehat{\xi}_{i \rightarrow 1}, \dots, \widehat{\xi}_{i \rightarrow M})^\top \right), \\
r_{\mathcal{H}}(\mathcal{H}) &= \prod_{n=1}^N r_z(\mathbf{Z}^{(n)}), \\
r_z(\mathbf{Z}^{(n)}) &= \frac{1}{C_{\mathbf{Z}^{(n)}}} \exp(\gamma_{\mathbf{Z}^{(n)}}), \\
\gamma_{\mathbf{Z}^{(n)}} &= \sum_{i,j,k=1}^K c_{i \rightarrow jk}^{\mathbf{Z}^{(n)}} \left\{ \Psi(\widehat{\phi}_{i \rightarrow jk}) - \Psi \left(\sum_{j'=1}^K \sum_{k'=1}^K \widehat{\phi}_{i \rightarrow j'k'} \right) \right\} \\
&\quad + \sum_{l=1}^L \sum_{i=1}^K \sum_{m=1}^M \bar{z}_i^{(n,l)} x_m^{(n,l)} \left\{ \Psi(\widehat{\xi}_{i \rightarrow m}) - \Psi \left(\sum_{m'=1}^M \widehat{\xi}_{i \rightarrow m'} \right) \right\},
\end{aligned} \tag{16.31}$$

where $C_{\mathbf{Z}^{(n)}} = \sum_{\mathbf{Z} \in T(X^{(n)})} \exp(\gamma_{\mathbf{Z}})$ is the normalizing constant and $T(X^{(n)})$ is the set of derivation sequences that generate $X^{(n)}$. After the substitution of Eq. (15.5), the free energy is given by

$$\begin{aligned}
F &= \sum_{i=1}^K \left\{ \log \left(\frac{\Gamma(\sum_{j,k=1}^K \widehat{\phi}_{i \rightarrow jk})}{\prod_{j,k=1}^K \Gamma(\widehat{\phi}_{i \rightarrow jk})} \right) \right. \\
&\quad + \sum_{j,k=1}^K (\widehat{\phi}_{i \rightarrow jk} - \phi) (\Psi(\widehat{\phi}_{i \rightarrow jk}) - \Psi(\sum_{j',k'=1}^K \widehat{\phi}_{i \rightarrow j'k'})) \\
&\quad + \log \left(\frac{\Gamma(\sum_{m=1}^M \widehat{\xi}_{i \rightarrow m})}{\prod_{m=1}^M \Gamma(\widehat{\xi}_{i \rightarrow m})} \right) + \sum_{m=1}^M (\widehat{\xi}_{i \rightarrow m} - \xi) (\Psi(\widehat{\xi}_{i \rightarrow m}) - \Psi(\sum_{m'=1}^M \widehat{\xi}_{i \rightarrow m'})) \Big\} \\
&\quad - K \log \left(\frac{\Gamma(K^2 \phi)}{(\Gamma(\phi))^{K^2}} \right) - K \log \left(\frac{\Gamma(M \xi)}{(\Gamma(\xi))^M} \right) - \sum_{n=1}^N \log C_{\mathbf{Z}^{(n)}}.
\end{aligned}$$

The variational parameters satisfy

$$\begin{aligned}
\widehat{\phi}_{i \rightarrow jk} &= \overline{N}_{i \rightarrow jk}^z + \phi, \\
\widehat{\xi}_{i \rightarrow m} &= \overline{N}_{i \rightarrow m}^x + \xi,
\end{aligned}$$

where

$$\begin{aligned}
\overline{N}_{i \rightarrow jk}^z &= \sum_{n=1}^N \sum_{l=1}^L \left\langle c_{i \rightarrow jk}^{\mathbf{Z}^{(n)}} \right\rangle_{r_z(\mathbf{Z}^{(n)})}, \\
\overline{N}_{i \rightarrow m}^x &= \sum_{n=1}^N \sum_{l=1}^L \left\langle \bar{z}_i^{(n,l)} \right\rangle_{r_z(\mathbf{Z}^{(n)})} x_m^{(n,l)}.
\end{aligned}$$

We assume the following condition.

Assumption 16.4 *The true distribution $q(X)$ has K_0 nonterminal symbols and M terminal symbols with parameter \mathbf{w}^* :*

$$q(X) = p(X|\mathbf{w}^*) = \sum_{Z \in T(X)} p(X, Z|\mathbf{w}^*). \quad (16.32)$$

The true parameters are

$$\begin{aligned}\mathbf{w}^* &= \{\{\mathbf{a}_i^*\}_{i=1}^{K_0}, \{\mathbf{b}_i^*\}_{i=1}^{K_0}\}, \\ \mathbf{a}_i^* &= \{a_{i \rightarrow jk}^*\}_{j,k=1}^{K_0} \quad (1 \leq i \leq K_0), \\ \mathbf{b}_i^* &= \{b_{i \rightarrow m}^*\}_{m=1}^M \quad (1 \leq i \leq K_0),\end{aligned}$$

which satisfy the constraints

$$a_{i \rightarrow ii}^* = 1 - \sum_{(j,k) \neq (i,i)} a_{i \rightarrow jk}^*, \quad b_{i \rightarrow M}^* = 1 - \sum_{m=1}^{M-1} b_{i \rightarrow m}^*,$$

respectively. Since PCFG has nontrivial nonidentifiability as in HMM (Ito et al., 1992), we assume that K_0 is the smallest number of nonterminal symbols under this parameterization. The statistical model given by Eq. (16.30) includes the true distribution, namely, the number of nonterminal symbols K satisfies the inequality $K_0 \leq K$.

Under this assumption, the next theorem evaluates the relative VB free energy. Here $S_N(\mathcal{D}) = -\frac{1}{N} \sum_{n=1}^N \log p(X^{(n)}|\mathbf{w}^*)$ is the empirical entropy of the true distribution (16.32).

Theorem 16.6 *The relative VB free energy of the PCFG model satisfies*

$$\tilde{F}^{\text{VB}}(\mathcal{D}) = F^{\text{VB}}(\mathcal{D}) - NS_N(\mathcal{D}) = \lambda'_{\text{PCFG}}^{\text{VB}} \log N + O_p(1),$$

where

$$\lambda'_{\text{PCFG}}^{\text{VB}} = \begin{cases} \frac{K_0(K_0^2-1)+K_0(M-1)}{2} + K_0(K^2 - K_0^2)\phi & \left(0 < \phi \leq \frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)}\right), \\ \frac{K(K^2-1)+K(M-1)}{2} & \left(\frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)} < \phi\right). \end{cases} \quad (16.33)$$

Proof Based on Lemma 15.1, similarly to the models in the previous sections, we evaluate $R = \text{KL}(r_w(\mathbf{w})||p(\mathbf{w}))$. It is expressed by expected sufficient statistics as

$$\begin{aligned}R &= \sum_{i=1}^K \left[\log \Gamma(\bar{N}_i^z + K^2\phi) - \bar{N}_i^z \Psi(\bar{N}_i^z + K^2\phi) \right. \\ &\quad \left. - \sum_{j,k=1}^K \left\{ \log \Gamma(\bar{N}_{i \rightarrow jk}^z + \phi) - \bar{N}_{i \rightarrow jk}^z \Psi(\bar{N}_{i \rightarrow jk}^z + \phi) \right\} \right]\end{aligned}$$

$$\begin{aligned}
& + \log \Gamma(\bar{N}_i^x + M\xi) - \bar{N}_i^x \Psi(\bar{N}_i^x + M\xi) \\
& - \sum_{m=1}^M \left\{ \log \Gamma(\bar{N}_{i \rightarrow m}^x + \xi) - \bar{N}_{i \rightarrow m}^x \Psi(\bar{N}_{i \rightarrow m}^x + \xi) \right\} \Big] + O_p(1).
\end{aligned}$$

Using the inequalities of the di-gamma and the log-gamma functions in Eqs. (15.8) and (15.9), we have

$$\begin{aligned}
R = & \sum_{i=1}^K \left[\left(K^2\phi - \frac{1}{2} \right) \log(\bar{N}_i^z + K^2\phi) - \sum_{j,k=1}^K \left(\phi - \frac{1}{2} \right) \log(\bar{N}_{i \rightarrow jk}^z + \phi) \right. \\
& \left. + \left(M\xi - \frac{1}{2} \right) \log(\bar{N}_i^x + M\xi) - \sum_{m=1}^M \left(\xi - \frac{1}{2} \right) \log(\bar{N}_{i \rightarrow m}^x + \xi) \right] + O_p(1). \tag{16.34}
\end{aligned}$$

We divide the sum over i , j , and k in Eq. (16.34) to the necessary K_0 and redundant $K - K_0$ terms. Moreover, we assume the trained model uses redundant l ($0 \leq l \leq K - K_0$) nonterminal symbols, i.e., it holds that $\bar{N}_i^z = \Theta_p(N)$:

$$\begin{aligned}
\frac{R}{\log N} = & \sum_{i=1}^{K_0} \left\{ \left(K^2\phi + \frac{M}{2} - 1 \right) - \sum_{j,k=1}^{K_0} \left(\phi - \frac{1}{2} \right) \right\} \\
& + \sum_{i=1}^{K_0} \left\{ \sum_{j,k=1}^{K_0} \left(\phi - \frac{1}{2} \right) - \sum_{j,k=1}^{K_0+l} \left(\phi - \frac{1}{2} \right) \right\} \\
& + \sum_{i=K_0+1}^{K_0+l} \left\{ \left(K^2\phi + \frac{M}{2} - 1 \right) - \sum_{j,k=1}^{K_0+l} \left(\phi - \frac{1}{2} \right) \right\} + O_p\left(\frac{1}{\log N}\right) \\
= & \left(K^2\phi + \frac{M}{2} - 1 \right) - K_0^2 \left(\phi - \frac{1}{2} \right) + g(l) + O_p\left(\frac{1}{\log N}\right), \tag{16.35}
\end{aligned}$$

where $g(l)$ is given by

$$g(l) = \left(K^2\phi + \frac{M}{2} - 1 \right) l - \left(\phi - \frac{1}{2} \right) \left\{ (K_0 + l)^3 - K_0^3 \right\}.$$

By Lemma 15.4, similarly to the HMM, we can evaluate the VB free energy by minimizing $g(l)$. The minimum of $g(l)$ is achieved by

$$\begin{cases} l = 0 & (0 < \phi \leq \frac{K_0^2 + KK_0 + K^2 + M - 2}{2(K_0^2 + KK_0)}), \\ l = K - K_0 & (\frac{K_0^2 + KK_0 + K^2 + M - 2}{2(K_0^2 + KK_0)} < \phi). \end{cases}$$

Putting this back into Eq. (16.35), we obtain the theorem. \square

16.4 Latent Dirichlet Allocation

In this section, we investigate the VB free energy of the latent Dirichlet allocation (LDA) introduced in Section 4.2.4. We also analyze the asymptotic behavior of MAP learning and partially Bayesian learning, which are often used alternatively to VB learning, and discuss similarities and dissimilarities between those learning algorithms.

We consider the following LDA model:

$$\begin{aligned} p(\mathbf{w}^{(n,m)}, \mathbf{z}^{(n,m)} | \boldsymbol{\Theta}, \mathbf{B}) &= p(\mathbf{w}^{(n,m)} | \mathbf{z}^{(n,m)}, \mathbf{B}) p(\mathbf{z}^{(n,m)} | \boldsymbol{\Theta}), \\ p(\mathbf{w}^{(n,m)} | \mathbf{z}^{(n,m)}, \mathbf{B}) &= \prod_{l=1}^L \prod_{h=1}^H (B_{l,h})^{w_l^{(n,m)} z_h^{(n,m)}}, \\ p(\mathbf{z}^{(n,m)} | \boldsymbol{\Theta}) &= \prod_{h=1}^H (\theta_{m,h})^{z_h^{(n,m)}}, \\ p(\boldsymbol{\Theta} | \alpha) &= \prod_{m=1}^M \text{Dirichlet}_H(\tilde{\boldsymbol{\theta}}_m; (\alpha, \dots, \alpha)^\top), \\ p(\mathbf{B} | \eta) &= \prod_{h=1}^H \text{Dirichlet}_L(\boldsymbol{\beta}_h; (\eta, \dots, \eta)^\top). \end{aligned}$$

Here we have assumed that the priors are symmetric and have hyperparameters $\alpha_1 = \dots = \alpha_H = \alpha > 0$, $\eta_1 = \dots = \eta_L = \eta > 0$, respectively.

Under the constraint, $r(\mathbf{w}, \mathcal{H}) = r_{\Theta, B}(\boldsymbol{\Theta}, \mathbf{B}) r_z(\{\{\mathbf{z}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M)$, the VB posteriors are given by

$$\begin{aligned} r_z(\{\{\mathbf{z}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M) &= \prod_{m=1}^M \prod_{n=1}^{N^{(m)}} \text{Multinomial}_{H,1}(\mathbf{z}^{(n,m)}; \widehat{\mathbf{z}}^{(n,m)}), \\ r_{\Theta, B}(\boldsymbol{\Theta}, \mathbf{B}) &= r_\Theta(\boldsymbol{\Theta}) r_B(\mathbf{B}), \\ r_\Theta(\boldsymbol{\Theta}) &= \prod_{m=1}^M \text{Dirichlet}(\tilde{\boldsymbol{\theta}}_m; \widehat{\boldsymbol{\alpha}}_m), \\ r_B(\mathbf{B}) &= \prod_{h=1}^H \text{Dirichlet}(\boldsymbol{\beta}_h; \widehat{\boldsymbol{\eta}}_h). \end{aligned}$$

The free energy is given by

$$\begin{aligned} F &= \sum_{m=1}^M \left(\log \left(\frac{\Gamma(\sum_{h=1}^H \widehat{\alpha}_{m,h})}{\prod_{h=1}^H \Gamma(\widehat{\alpha}_{m,h})} \right) - \log \left(\frac{\Gamma(H\alpha)}{\Gamma(\alpha)^H} \right) \right) \\ &\quad + \sum_{h=1}^H \left(\log \left(\frac{\Gamma(\sum_{l=1}^L \widehat{\eta}_{l,h})}{\prod_{l=1}^L \Gamma(\widehat{\eta}_{l,h})} \right) - \log \left(\frac{\Gamma(L\eta)}{\Gamma(\eta)^L} \right) \right) \end{aligned}$$

$$\begin{aligned}
& + \sum_{m=1}^M \sum_{h=1}^H \left(\widehat{\alpha}_{m,h} - (\bar{N}_h^{(m)} + \alpha) \right) \left(\Psi(\widehat{\alpha}_{m,h}) - \Psi(\sum_{h'=1}^H \widehat{\alpha}_{m,h'}) \right) \\
& + \sum_{h=1}^H \sum_{l=1}^L \left(\widehat{\eta}_{l,h} - (\bar{W}_{l,h} + \eta_l) \right) \left(\Psi(\widehat{\eta}_{l,h}) - \Psi(\sum_{l'=1}^L \widehat{\eta}_{l',h}) \right) \\
& + \sum_{m=1}^M \sum_{n=1}^{N^{(m)}} \sum_{h=1}^H \bar{z}_h^{(n,m)} \log \bar{z}_h^{(n,m)},
\end{aligned}$$

where

$$\bar{N}_h^{(m)} = \sum_{n=1}^{N^{(m)}} \langle z_h^{(n,m)} \rangle_{r_z(\{z^{(n,m)}\}_{n=1}^{N^{(m)}})_{m=1}^M},$$

$$\bar{W}_{l,h} = \sum_{m=1}^M \sum_{n=1}^{N^{(m)}} w_l^{(n,m)} \langle z_h^{(n,m)} \rangle_{r_z(\{z^{(n,m)}\}_{n=1}^{N^{(m)}})_{m=1}^M}},$$

for the observed data $\mathcal{D} = \{\{w^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M$. The variational parameters satisfy

$$\widehat{\alpha}_{m,h} = \bar{N}_h^{(m)} + \alpha, \quad (16.36)$$

$$\widehat{\eta}_{l,h} = \bar{W}_{l,h} + \eta_l, \quad (16.37)$$

$$\bar{z}_h^{(n,m)} = \frac{\bar{z}_h^{(n,m)}}{\sum_{h'=1}^H \bar{z}_{h'}^{(n,m)}}$$

for

$$\begin{aligned}
\bar{z}_h^{(n,m)} &= \exp \left(\left\{ \Psi(\widehat{\alpha}_{m,h}) - \Psi \left(\sum_{h'=1}^H \widehat{\alpha}_{m,h'} \right) \right\} \right. \\
&\quad \left. + \sum_{l=1}^L w_l^{(n,m)} \left\{ \Psi(\widehat{\eta}_{l,h}) - \Psi \left(\sum_{l'=1}^L \widehat{\eta}_{l',h} \right) \right\} \right).
\end{aligned}$$

Based on Lemma 15.1, we decompose the free energy as follows:

$$F = R + Q, \quad (16.38)$$

where

$$\begin{aligned}
R &= \text{KL}(r_\Theta(\boldsymbol{\theta}) r_B(\mathbf{B}) || p(\boldsymbol{\theta}|\alpha) p(\mathbf{B}|\eta)) \\
&= \sum_{m=1}^M \left(\log \frac{\Gamma(\sum_{h=1}^H \widehat{\alpha}_{m,h})}{\prod_{h=1}^H \Gamma(\widehat{\alpha}_{m,h})} \frac{\Gamma(\alpha)^H}{\Gamma(H\alpha)} + \sum_{h=1}^H (\widehat{\alpha}_{m,h} - \alpha) \left(\Psi(\widehat{\alpha}_{m,h}) - \Psi(\sum_{h'=1}^H \widehat{\alpha}_{m,h'}) \right) \right) \\
&\quad + \sum_{h=1}^H \left(\log \frac{\Gamma(\sum_{l=1}^L \widehat{\eta}_{l,h})}{\prod_{l=1}^L \Gamma(\widehat{\eta}_{l,h})} \frac{\Gamma(\eta_l)^L}{\Gamma(L\eta_l)} + \sum_{l=1}^L (\widehat{\eta}_{l,h} - \eta_l) \left(\Psi(\widehat{\eta}_{l,h}) - \Psi(\sum_{l'=1}^L \widehat{\eta}_{l',h}) \right) \right),
\end{aligned} \quad (16.39)$$

$$Q = -\log C_{\mathcal{H}}$$

$$= - \sum_{m=1}^M N^{(m)} \sum_{l=1}^L V_{l,m} \log \left(\sum_{h=1}^H \frac{\exp(\Psi(\widehat{\alpha}_{m,h}))}{\exp(\Psi(\sum_{h'=1}^H \widehat{\alpha}_{m,h'}))} \frac{\exp(\Psi(\widehat{\eta}_{l,h}))}{\exp(\Psi(\sum_{l'=1}^L \widehat{\eta}_{l',h}))} \right). \quad (16.40)$$

Here, $\mathbf{V} \in \mathbb{R}^{L \times M}$ is the empirical word distribution matrix with its entries given by $V_{l,m} = \frac{1}{N^{(m)}} \sum_{n=1}^{N^{(m)}} w_l^{(n,m)}$.

16.4.1 Asymptotic Analysis of VB Learning

Here we analyze the VB free energy of LDA in the asymptotic limit when $N \equiv \min_m N^{(m)} \rightarrow \infty$ (Nakajima et al., 2014). Unlike the analyses for the latent variable models in the previous sections, we do not assume $L, M \ll N$, but $1 \ll L, M, N$ at this point. This amounts to considering the asymptotic limit when $L, M, N \rightarrow \infty$ with a fixed mutual ratio, or equivalently, assuming $L, M \sim O(N)$.

We assume the following condition on the true distribution.

Assumption 16.5 *The word distribution matrix \mathbf{V} is a sample from the multinomial distribution with the true parameter $\mathbf{U}^* \in \mathbb{R}^{L \times M}$ whose rank is $H^* \sim O(1)$, i.e., $\mathbf{U}^* = \mathbf{B}^* \boldsymbol{\Theta}^{*\top}$ where $\boldsymbol{\Theta}^* \in \mathbb{R}^{M \times H^*}$ and $\mathbf{B}^* \in \mathbb{R}^{L \times H^*}$.² The number of topics of the model H is set to $H = \min(L, M)$ (i.e., the matrix $\mathbf{B}\boldsymbol{\Theta}^\top$ can express any multinomial distribution).*

The stationary conditions, Eqs. (16.36) and (16.37), lead to the following lemma:

Lemma 16.7 *Let $\widehat{\mathbf{B}\boldsymbol{\Theta}}^\top = \langle \mathbf{B}\boldsymbol{\Theta}^\top \rangle_{r_{\boldsymbol{\Theta}, \mathbf{B}}(\boldsymbol{\Theta}, \mathbf{B})}$. Then it holds that*

$$\left\langle (\mathbf{B}\boldsymbol{\Theta}^\top - \widehat{\mathbf{B}\boldsymbol{\Theta}}^\top)^2 \right\rangle_{r_{\boldsymbol{\Theta}, \mathbf{B}}(\boldsymbol{\Theta}, \mathbf{B})} = O_p(N^{-2}), \quad (16.41)$$

$$Q = - \sum_{m=1}^M N^{(m)} \sum_{l=1}^L V_{l,m} \log \langle \widehat{\mathbf{B}\boldsymbol{\Theta}}^\top \rangle_{l,m} + O_p(M). \quad (16.42)$$

Proof For the Dirichlet distribution $p(\mathbf{a}|\check{\mathbf{a}}) \propto \prod_{h=1}^H a_h^{\check{a}_h - 1}$, the mean and the variance are given as follows:

$$\widehat{a}_h = \langle a_h \rangle_{p(\mathbf{a}|\check{\mathbf{a}})} = \frac{\check{a}_h}{\check{a}_0}, \quad \langle (a_h - \widehat{a}_h)^2 \rangle_{p(\mathbf{a}|\check{\mathbf{a}})} = \frac{\check{a}_h(\check{a}_0 - \check{a}_h)}{\check{a}_0^2(\check{a}_0 + 1)},$$

where $\check{a}_0 = \sum_{h=1}^H \check{a}_h$.

² More precisely, $\mathbf{U}^* = \mathbf{B}^* \boldsymbol{\Theta}^{*\top} + O(N^{-1})$ is sufficient.

For fixed N, R , defined by Eq. (16.39), diverges to $+\infty$ if $\widehat{\alpha}_{m,h} \rightarrow +0$ for any (m, h) or $\widehat{\eta}_{l,h} \rightarrow +0$ for any (l, h) . Therefore, the global minimizer of the free energy (16.38) is in the interior of the domain, where the free energy is differentiable. Consequently, the global minimizer is a stationary point. The stationary conditions (16.36) and (16.37) imply that

$$\widehat{\alpha}_{m,h} \geq \alpha, \quad \widehat{\eta}_{l,h} \geq \eta, \quad (16.43)$$

$$\sum_{h=1}^H \widehat{\alpha}_{m,h} = \sum_{h=1}^H \alpha + N^{(m)}, \quad \sum_{l=1}^L \widehat{\eta}_{l,h} = \sum_{l=1}^L \eta + \sum_{m=1}^M (\widehat{\alpha}_{m,h} - \alpha). \quad (16.44)$$

Therefore, we have

$$\langle (\Theta_{m,h} - \widehat{\Theta}_{m,h})^2 \rangle_{r_{\Theta}(\boldsymbol{\theta})} = O_p(N^{-2}) \quad \text{for all } (m, h), \quad (16.45)$$

$$\left(\max_m \widehat{\Theta}_{m,h} \right)^2 \langle (B_{l,h} - \widehat{B}_{l,h})^2 \rangle_{r_B(\boldsymbol{B})} = O_p(N^{-2}) \quad \text{for all } (l, h), \quad (16.46)$$

which leads to Eq. (16.41).

By using Eq. (15.8), Q is bounded as follows:

$$\underline{Q} \leq Q \leq \overline{Q},$$

where

$$\begin{aligned} \overline{Q} &= - \sum_{m=1}^M N^{(m)} \sum_{l=1}^L V_{l,m} \log \left(\sum_{h=1}^H \frac{\widehat{\alpha}_{m,h}}{\sum_{h'=1}^H \widehat{\alpha}_{m,h'}} \frac{\widehat{\eta}_{l,h}}{\sum_{l'=1}^L \widehat{\eta}_{l',h}} \frac{\exp\left(-\frac{1}{\widehat{\alpha}_{m,h}}\right)}{\exp\left(-\frac{1}{2 \sum_{h'=1}^H \widehat{\alpha}_{m,h'}}\right)} \frac{\exp\left(-\frac{1}{\widehat{\eta}_{l,h}}\right)}{\exp\left(-\frac{1}{2 \sum_{l'=1}^L \widehat{\eta}_{l',h}}\right)} \right), \\ \underline{Q} &= - \sum_{m=1}^M N^{(m)} \sum_{l=1}^L V_{l,m} \log \left(\sum_{h=1}^H \frac{\widehat{\alpha}_{m,h}}{\sum_{h'=1}^H \widehat{\alpha}_{m,h'}} \frac{\widehat{\eta}_{l,h}}{\sum_{l'=1}^L \widehat{\eta}_{l',h}} \frac{\exp\left(-\frac{1}{2\widehat{\alpha}_{m,h}}\right)}{\exp\left(-\frac{1}{\sum_{h'=1}^H \widehat{\alpha}_{m,h'}}\right)} \frac{\exp\left(-\frac{1}{2\widehat{\eta}_{l,h}}\right)}{\exp\left(-\frac{1}{\sum_{l'=1}^L \widehat{\eta}_{l',h}}\right)} \right). \end{aligned}$$

Using Eqs. (16.45) and (16.46), we have Eq. (16.42), which completes the proof of Lemma 16.7. \square

Eq. (16.41) implies the convergence of the posterior. Let $\boldsymbol{u}_m^* = \boldsymbol{B}^*(\widehat{\boldsymbol{\theta}}_m^*)^\top$ be the true probability mass function for the m th document and $\widehat{\boldsymbol{u}}_m = \widehat{\boldsymbol{B}}(\widehat{\boldsymbol{\theta}}_m)^\top$ be its predictive probability. Define a measure of how far the predictive distributions are from the true distributions by

$$\widehat{J} = \sum_{m=1}^M \frac{N^{(m)}}{N} \text{KL}(\boldsymbol{u}_m^* || \widehat{\boldsymbol{u}}_m). \quad (16.47)$$

Then, by the same arguments as the proof of Lemma 15.4, Eq. (16.42) leads to the following lemma:

Lemma 16.8 Q is minimized when $\widehat{J} = O_p(1/N)$, and it holds that

$$Q = NS_N(\mathcal{D}) + O_p(\widehat{J}N + LM), \quad \text{where}$$

$$S_N(\mathcal{D}) = -\frac{1}{N} \log p(\mathcal{D} | \boldsymbol{\Theta}^*, \mathbf{B}^*) = -\sum_{m=1}^M \frac{N^{(m)}}{N} \sum_{l=1}^L V_{l,m} \log(\mathbf{B}^* \boldsymbol{\Theta}^*)_{l,m}.$$

Lemma 16.8 simply states that Q/N converges to the empirical entropy $S_N(\mathcal{D})$ of the true distribution if and only if the predictive distribution converges to the true distribution (i.e., $\widehat{J} = O_p(1/N)$).

Let $\widehat{H} = \sum_{h=1}^H \theta(\frac{1}{M} \sum_{m=1}^M \widehat{\Theta}_{m,h} \sim O_p(1))$ be the number of topics used in the whole corpus, $\widehat{M}^{(h)} = \sum_{m=1}^M \theta(\widehat{\Theta}_{m,h} \sim O_p(1))$ be the number of documents that contain the h th topic, and $\widehat{L}^{(h)} = \sum_{l=1}^L \theta(\widehat{B}_{l,h} \sim O_p(1))$ be the number of words of which the h th topic consist. We have the following lemma:

Lemma 16.9 R is written as follows:

$$\begin{aligned} R = & \left\{ M \left(H\alpha - \frac{1}{2} \right) + \widehat{H} \left(L\eta - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \left(\widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) + \widehat{L}^{(h)} \left(\eta - \frac{1}{2} \right) \right) \right\} \log N \\ & + (H - \widehat{H}) \left(L\eta - \frac{1}{2} \right) \log L + O_p(H(M + L)). \end{aligned} \quad (16.48)$$

Proof By using the bounds (15.8) and (15.9), R can be bounded as

$$\underline{R} \leq R \leq \bar{R}, \quad (16.49)$$

where

$$\begin{aligned} \underline{R} = & -\sum_{m=1}^M \log \left(\frac{\Gamma(H\alpha)}{\Gamma(\alpha)^H} \right) - \sum_{h=1}^H \log \left(\frac{\Gamma(L\eta)}{\Gamma(\eta)^L} \right) - \frac{M(H-1) + H(L-1)}{2} \log(2\pi) \\ & + \sum_{m=1}^M \left\{ \left(H\alpha - \frac{1}{2} \right) \log \sum_{h=1}^H \widehat{\alpha}_{m,h} - \sum_{h=1}^H \left(\alpha - \frac{1}{2} \right) \log \widehat{\alpha}_{m,h} \right\} \\ & + \sum_{h=1}^H \left\{ \left(L\eta - \frac{1}{2} \right) \log \sum_{l=1}^L \widehat{\eta}_{l,h} - \sum_{l=1}^L \left(\eta - \frac{1}{2} \right) \log \widehat{\eta}_{l,h} \right\} \\ & + \sum_{m=1}^M \left\{ -\sum_{h=1}^H \frac{1}{12\widehat{\alpha}_{m,h}} - \sum_{h=1}^H (\widehat{\alpha}_{m,h} - \alpha) \left(\frac{1}{\widehat{\alpha}_{m,h}} - \frac{1}{2 \sum_{h'=1}^H \widehat{\alpha}_{m,h'}} \right) \right\} \\ & + \sum_{h=1}^H \left\{ -\sum_{l=1}^L \frac{1}{12\widehat{\eta}_{l,h}} - \sum_{l=1}^L (\widehat{\eta}_{l,h} - \eta) \left(\frac{1}{\widehat{\eta}_{l,h}} - \frac{1}{2 \sum_{l'=1}^L \widehat{\eta}_{l',h}} \right) \right\}, \end{aligned} \quad (16.50)$$

$$\begin{aligned}
\bar{R} = & - \sum_{m=1}^M \log \left(\frac{\Gamma(H\alpha)}{\Gamma(\alpha)^H} \right) - \sum_{h=1}^H \log \left(\frac{\Gamma(L\eta)^L}{\Gamma(\eta)} \right) - \frac{M(H-1) + H(L-1)}{2} \log(2\pi) \\
& + \sum_{m=1}^M \left\{ \left(H\alpha - \frac{1}{2} \right) \log \sum_{h=1}^H \widehat{\alpha}_{m,h} - \sum_{h=1}^H \left(\alpha - \frac{1}{2} \right) \log \widehat{\alpha}_{m,h} \right\} \\
& + \sum_{h=1}^H \left\{ \left(L\eta - \frac{1}{2} \right) \log \sum_{l=1}^L \widehat{\eta}_{l,h} - \sum_{l=1}^L \left(\eta - \frac{1}{2} \right) \log \widehat{\eta}_{l,h} \right\} \\
& + \sum_{m=1}^M \left\{ \frac{1}{12 \sum_{h=1}^H \widehat{\alpha}_{m,h}} - \sum_{h=1}^H (\widehat{\alpha}_{m,h} - \alpha) \left(\frac{1}{2\widehat{\alpha}_{m,h}} - \frac{1}{\sum_{h'=1}^H \widehat{\alpha}_{m,h'}} \right) \right\} \\
& + \sum_{h=1}^H \left\{ \frac{1}{12 \sum_{l=1}^L \widehat{\eta}_{l,h}} - \sum_{l=1}^L (\widehat{\eta}_{l,h} - \eta) \left(\frac{1}{2\widehat{\eta}_{l,h}} - \frac{1}{\sum_{l'=1}^L \widehat{\eta}_{l',h}} \right) \right\}. \tag{16.51}
\end{aligned}$$

Eqs. (16.43) and (16.44) imply that

$$\begin{aligned}
R = & \sum_{m=1}^M \left\{ \left(H\alpha - \frac{1}{2} \right) \log \sum_{h=1}^H \widehat{\alpha}_{m,h} - \sum_{h=1}^H \left(\alpha - \frac{1}{2} \right) \log \widehat{\alpha}_{m,h} \right\} \\
& + \sum_{h=1}^H \left\{ \left(L\eta - \frac{1}{2} \right) \log \sum_{l=1}^L \widehat{\eta}_{l,h} - \sum_{l=1}^L \left(\eta - \frac{1}{2} \right) \log \widehat{\eta}_{l,h} \right\} + O_p(H(M+L)),
\end{aligned}$$

which leads to Eq. (16.48). This completes the proof of Lemma 16.9. \square

Since we assumed that the true matrices $\boldsymbol{\Theta}^*$ and \mathbf{B}^* are of the rank of H^* , $\widehat{H} = H^* \sim O(1)$ is sufficient for the VB posterior to converge to the *true* distribution. However, \widehat{H} can be much larger than H^* with $\langle \mathbf{B}\boldsymbol{\Theta}^\top \rangle_{r_{\Theta,B}(\boldsymbol{\Theta}, \mathbf{B})}$ unchanged because of the nonidentifiability of matrix factorization—duplicating topics with divided weights, for example, does not change the distribution.

Let

$$\widetilde{F}^{\text{VB}}(\mathcal{D}) = F^{\text{VB}}(\mathcal{D}) - NS_N(\mathcal{D}) \tag{16.52}$$

be the relative free energy. Based on Lemmas 16.8 and 16.9, we obtain the following theorem:

Theorem 16.10 *In the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$, it holds that $\widehat{J} = O_p(1/N)$, and*

$$\widetilde{F}^{\text{VB}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{VB}} \log N + O_p(1),$$

where

$$\lambda'_{\text{LDA}}^{\text{VB}} = \left\{ M \left(H\alpha - \frac{1}{2} \right) + \widehat{H} \left(L\eta - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \left(\widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) + \widehat{L}^{(h)} \left(\eta - \frac{1}{2} \right) \right) \right\}.$$

In the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$, it holds that $\widehat{J} = o_p(\log N)$, and

$$\widetilde{F}^{\text{VB}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{VB}} \log N + o_p(N \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{VB}} = \left\{ M \left(H\alpha - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) \right\}.$$

In the limit when $N, L \rightarrow \infty$ with $\frac{L}{N}, M \sim O(1)$, it holds that $\widehat{J} = o_p(\log N)$, and

$$\widetilde{F}^{\text{VB}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{VB}} \log N + o_p(N \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{VB}} = HL\eta.$$

In the limit when $N, L, M \rightarrow \infty$ with $\frac{L}{N}, \frac{M}{N} \sim O(1)$, it holds that $\widehat{J} = o_p(N \log N)$, and

$$\widetilde{F}^{\text{VB}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{VB}} \log N + o_p(N^2 \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{VB}} = H(M\alpha + L\eta).$$

Proof Lemmas 16.8 and 16.9 imply that the relative free energy can be written as follows:

$$\begin{aligned} \widetilde{F}^{\text{VB}}(\mathcal{D}) &= \left\{ M \left(H\alpha - \frac{1}{2} \right) + \widehat{H} \left(L\eta - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \left(\widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) + \widehat{L}^{(h)} \left(\eta - \frac{1}{2} \right) \right) \right\} \log N \\ &\quad + (H - \widehat{H}) \left(L\eta - \frac{1}{2} \right) \log L + O_p(\widehat{J}N + LM). \end{aligned} \tag{16.53}$$

In the following subsection, we investigate the leading term of the relative free energy (16.53) in different asymptotic limits.

In the Limit When $N \rightarrow \infty$ with $L, M \sim O(1)$

In this case, the minimizer should satisfy

$$\widehat{J} = O_p \left(\frac{1}{N} \right) \tag{16.54}$$

and the leading term of the relative free energy (16.52) is of the order of $O_p(\log N)$ as follows:

$$\begin{aligned} \tilde{F}^{\text{VB}}(\mathcal{D}) &= \left\{ M \left(H\alpha - \frac{1}{2} \right) + \widehat{H} \left(L\eta - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \left(\widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) + \widehat{L}^{(h)} \left(\eta - \frac{1}{2} \right) \right) \right\} \log N \\ &\quad + O_p(1). \end{aligned}$$

Note that Eq. (16.54) implies the consistency of the VB posterior.

In the Limit When $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$

In this case,

$$\widehat{J} = o_p(\log N), \quad (16.55)$$

making the leading term of the relative free energy of the order of $O_p(N \log N)$ as follows:

$$\tilde{F}^{\text{VB}}(\mathcal{D}) = \left\{ M \left(H\alpha - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) \right\} \log N + o_p(N \log N).$$

Eq. (16.55) implies that the VB posterior is not necessarily consistent.

In the Limit When $N, L \rightarrow \infty$ with $\frac{L}{N}, M \sim O(1)$

In this case, Eq. (16.55) holds, and the leading term of the relative free energy is of the order of $O_p(N \log N)$ as follows:

$$\tilde{F}^{\text{VB}}(\mathcal{D}) = HL\eta \log N + o_p(N \log N).$$

In the Limit When $N, L, M \rightarrow \infty$ with $\frac{L}{N}, \frac{M}{N} \sim O(1)$

In this case,

$$\widehat{J} = o_p(N \log N), \quad (16.56)$$

and the leading term of the relative free energy is of the order of $O_p(N^2 \log N)$ as follows:

$$\tilde{F}^{\text{VB}}(\mathcal{D}) = H(M\alpha + L\eta) \log N + o_p(N^2 \log N).$$

This completes the proof of Theorem 16.10. \square

Since Eq. (16.41) was shown to hold, the predictive distribution converges to the true distribution if $\widehat{J} = O_p(1/N)$. Accordingly, Theorem 16.10 states that the consistency holds in the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$.

Theorem 16.10 also implies that, in the asymptotic limits with small $L \sim O(1)$, the leading term depends on \widehat{H} , meaning that it dominates the topic sparsity of the VB solution. We have the following corollary:

Corollary 16.11 *Let $M^{*(h)} = \sum_{m=1}^M \theta(\Theta_{m,h}^* \sim O(1))$ and $L^{*(h)} = \sum_{l=1}^L \theta(B_{l,h}^* \sim O(1))$. Consider the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$. When $0 < \eta \leq \frac{1}{2L}$, the VB solution is sparse (i.e., $\widehat{H} \ll H = \min(L, M)$) if $\alpha < \frac{1}{2} - \frac{\frac{1}{2} - L\eta}{\min_h M^{*(h)}}$, and dense (i.e., $\widehat{H} \approx H$) if $\alpha > \frac{1}{2} - \frac{\frac{1}{2} - L\eta}{\min_h M^{*(h)}}$. When $\frac{1}{2L} < \eta \leq \frac{1}{2}$, the VB solution is sparse if $\alpha < \frac{1}{2} + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}$, and dense if $\alpha > \frac{1}{2} + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}$. When $\eta > \frac{1}{2}$, the VB solution is sparse if $\alpha < \frac{1}{2} + \frac{L-1}{2\max_h M^{*(h)}}$, and dense if $\alpha > \frac{1}{2} + \frac{L\eta - \frac{1}{2}}{\min_h M^{*(h)}}$. In the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$, the VB solution is sparse if $\alpha < \frac{1}{2}$, and dense if $\alpha > \frac{1}{2}$.*

Proof From the compact representation when $\widehat{H} = H^*, \widehat{M}^{(h)} = M^{*(h)}$, and $\widehat{L}^{(h)} = L^{*(h)}$, we can decompose a singular component into two, keeping $\widehat{B}\widehat{\Theta}^\top$ unchanged, so that

$$\widehat{H} \rightarrow \widehat{H} + 1, \quad (16.57)$$

$$\sum_{h=1}^H \widehat{M}^{(h)} \rightarrow \sum_{h=1}^{H^*} \widehat{M}^{(h)} + \Delta M \quad \text{for } \min_h M^{*(h)} \leq \Delta M \leq \max_h M^{*(h)}, \quad (16.58)$$

$$\sum_{h=1}^H \widehat{L}^{(h)} \rightarrow \sum_{h=1}^{H^*} \widehat{L}^{(h)} + \Delta L \quad \text{for } 0 \leq \Delta L \leq \max_h L^{*(h)}. \quad (16.59)$$

Here the lower-bound for ΔM in Eq. (16.58) corresponds to the case that the least frequent topic is chosen to be decomposed, while the upper-bound to the case that the most frequent topic is chosen. The lower-bound for ΔL in Eq. (16.59) corresponds to the decomposition such that some of the word-occurrences are moved to a new topic, while the upper-bound to the decomposition such that the topic with the widest vocabulary is copied to a new topic. Note that the bounds both for ΔM and ΔL are not always achievable simultaneously, when we choose one topic to decompose.

In the following subsection, we investigate the relation between the sparsity of the solution and the hyperparameter setting in different asymptotic limits.

In the Limit When $N \rightarrow \infty$ with $L, M \sim O(1)$

The coefficient of the leading term of the free energy is

$$\chi_{\text{LDA}}^{\text{VB}} = M \left(H\alpha - \frac{1}{2} \right) + \sum_{h=1}^{\widehat{H}} \left(L\eta - \frac{1}{2} - \widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) - \widehat{L}^{(h)} \left(\eta - \frac{1}{2} \right) \right). \quad (16.60)$$

Note that the solution is sparse if Eq. (16.60) is increasing with respect to \widehat{H} , and dense if it is decreasing. Eqs. (16.57) through (16.59) imply the following:

- (I) When $0 < \eta \leq \frac{1}{2L}$ and $\alpha \leq \frac{1}{2}$, the solution is sparse if

$$\begin{aligned} L\eta - \frac{1}{2} - \min_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) &> 0, \text{ or equivalently,} \\ \alpha &< \frac{1}{2} - \frac{1}{\min_h M^{*(h)}} \left(\frac{1}{2} - L\eta \right), \end{aligned}$$

and dense if

$$\alpha > \frac{1}{2} - \frac{1}{\min_h M^{*(h)}} \left(\frac{1}{2} - L\eta \right).$$

- (II) When $0 < \eta \leq \frac{1}{2L}$ and $\alpha > \frac{1}{2}$, the solution is sparse if

$$\begin{aligned} L\eta - \frac{1}{2} - \max_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) &> 0, \text{ or equivalently,} \\ \alpha &< \frac{1}{2} - \frac{1}{\max_h M^{*(h)}} \left(\frac{1}{2} - L\eta \right), \end{aligned}$$

and dense if

$$\alpha > \frac{1}{2} - \frac{1}{\max_h M^{*(h)}} \left(\frac{1}{2} - L\eta \right).$$

Therefore, the solution is always dense in this case.

- (III) When $\frac{1}{2L} < \eta \leq \frac{1}{2}$ and $\alpha < \frac{1}{2}$, the solution is sparse if

$$\begin{aligned} L\eta - \frac{1}{2} - \min_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) &> 0, \text{ or equivalently,} \\ \alpha &< \frac{1}{2} + \frac{1}{\min_h M^{*(h)}} \left(L\eta - \frac{1}{2} \right), \end{aligned}$$

and dense if

$$\alpha > \frac{1}{2} + \frac{1}{\min_h M^{*(h)}} \left(L\eta - \frac{1}{2} \right).$$

Therefore, the solution is always sparse in this case.

- (IV) When $\frac{1}{2L} < \eta \leq \frac{1}{2}$ and $\alpha \geq \frac{1}{2}$, the solution is sparse if

$$\begin{aligned} L\eta - \frac{1}{2} - \max_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) &> 0, \text{ or equivalently,} \\ \alpha &< \frac{1}{2} + \frac{1}{\max_h M^{*(h)}} \left(L\eta - \frac{1}{2} \right), \end{aligned}$$

and dense if

$$\alpha > \frac{1}{2} + \frac{1}{\max_h M^{*(h)}} \left(L\eta - \frac{1}{2} \right).$$

(V) When $\eta > \frac{1}{2}$ and $\alpha < \frac{1}{2}$, the solution is sparse if

$$L\eta - \frac{1}{2} - \max_h \left(M^{*(h)} \left(\alpha - \frac{1}{2} \right) + L^{*(h)} \left(\eta - \frac{1}{2} \right) \right) > 0, \quad (16.61)$$

and dense if

$$L\eta - \frac{1}{2} - \max_h \left(M^{*(h)} \left(\alpha - \frac{1}{2} \right) + L^{*(h)} \left(\eta - \frac{1}{2} \right) \right) < 0. \quad (16.62)$$

Therefore, the solution is sparse if

$$\begin{aligned} L\eta - \frac{1}{2} - \min_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) - \max_h L^{*(h)} \left(\eta - \frac{1}{2} \right) &> 0, \text{ or equivalently,} \\ \alpha &< \frac{1}{2} + \frac{1}{\min_h M^{*(h)}} \left(L\eta - \frac{1}{2} - \max_h L^{*(h)} \left(\eta - \frac{1}{2} \right) \right), \end{aligned}$$

and dense if

$$\begin{aligned} L\eta - \frac{1}{2} - \max_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) - \max_h L^{*(h)} \left(\eta - \frac{1}{2} \right) &< 0, \text{ or equivalently,} \\ \alpha &> \frac{1}{2} + \frac{1}{\max_h M^{*(h)}} \left(L\eta - \frac{1}{2} - \max_h L^{*(h)} \left(\eta - \frac{1}{2} \right) \right). \end{aligned}$$

Therefore, the solution is always sparse in this case.

(VI) When $\eta > \frac{1}{2}$ and $\alpha \geq \frac{1}{2}$, the solution is sparse if Eq. (16.61) holds, and dense if Eq. (16.62) holds. Therefore, the solution is sparse if

$$\begin{aligned} L\eta - \frac{1}{2} - \max_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) - \max_h L^{*(h)} \left(\eta - \frac{1}{2} \right) &> 0, \text{ or equivalently,} \\ \alpha &< \frac{1}{2} + \frac{1}{\max_h M^{*(h)}} \left(L\eta - \frac{1}{2} - \max_h L^{*(h)} \left(\eta - \frac{1}{2} \right) \right), \end{aligned}$$

and dense if

$$\begin{aligned} L\eta - \frac{1}{2} - \min_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) - \max_h L^{*(h)} \left(\eta - \frac{1}{2} \right) &< 0, \text{ or equivalently,} \\ \alpha &> \frac{1}{2} + \frac{1}{\min_h M^{*(h)}} \left(L\eta - \frac{1}{2} - \max_h L^{*(h)} \left(\eta - \frac{1}{2} \right) \right). \end{aligned}$$

Thus, we can conclude that, in this case, the solution is sparse if

$$\alpha < \frac{1}{2} + \frac{L - 1}{2 \max_h M^{*(h)}},$$

and dense if

$$\alpha > \frac{1}{2} + \frac{L\eta - \frac{1}{2}}{\min_h M^{*(h)}}.$$

Summarizing the preceding, we have the following lemma:

Lemma 16.12 *When $0 < \eta \leq \frac{1}{2L}$, the solution is sparse if $\alpha < \frac{1}{2} - \frac{\frac{1}{2} - L\eta}{\min_h M^{*(h)}}$, and dense if $\alpha > \frac{1}{2} - \frac{\frac{1}{2} - L\eta}{\min_h M^{*(h)}}$. When $\frac{1}{2L} < \eta \leq \frac{1}{2}$, the solution is sparse if $\alpha < \frac{1}{2} + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}$, and dense if $\alpha > \frac{1}{2} + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}$. When $\eta > \frac{1}{2}$, the solution is sparse if $\alpha < \frac{1}{2} + \frac{L-1}{2\max_h M^{*(h)}}$, and dense if $\alpha > \frac{1}{2} + \frac{L\eta - \frac{1}{2}}{\min_h M^{*(h)}}$.*

In the Limit When $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$

The coefficient of the leading term of the free energy is given by

$$\lambda'_{\text{LDA}}^{\text{VB}} = M \left(H\alpha - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right). \quad (16.63)$$

Although the predictive distribution does not necessarily converge to the true distribution, we can investigate the sparsity of the solution by considering the duplication rules (16.57) through (16.59) that keep $\widehat{\mathbf{B}}\boldsymbol{\theta}^\top$ unchanged. It is clear that Eq. (16.63) is increasing with respect to \widehat{H} if $\alpha < \frac{1}{2}$, and decreasing if $\alpha > \frac{1}{2}$. Combing this result with Lemma 16.12 completes the proof of Corollary 16.11. \square

In the case when $L, M \ll N$ and in the case when $L \ll M, N$, Corollary 16.11 provides information on the sparsity of the VB solution, which will be compared with other methods in Section 16.4.2. On the other hand, although we have successfully derived the leading term of the free energy also in the case when $M \ll L, N$ and in the case when $1 \ll L, M, N$, it unfortunately provides no information on sparsity of the solution.

16.4.2 Asymptotic Analysis of MAP Learning and Partially Bayesian Learning

For training the LDA model, MAP learning and partially Bayesian (PB) learning (see Section 2.2.2), where $\boldsymbol{\Theta}$ and/or \mathbf{B} are point-estimated, are also popular choices. Although the differences in update equations is small, it can affect the asymptotic behavior. In this subsection, we aim to clarify the difference in the asymptotic behavior.

MAP learning, PB-A learning, PB-B learning, and VB learning, respectively, solve the following problem:

$$\min_r F,$$

s.t.

$$\begin{cases} r_{\Theta, B}(\boldsymbol{\Theta}, \mathbf{B}) = \delta(\boldsymbol{\Theta}; \widehat{\boldsymbol{\Theta}}) \delta(\mathbf{B}; \widehat{\mathbf{B}}) & \text{(for MAP learning),} \\ r_{\Theta, B}(\boldsymbol{\Theta}, \mathbf{B}) = r_{\Theta}(\boldsymbol{\Theta}) \delta(\mathbf{B}; \widehat{\mathbf{B}}) & \text{(for PB-A learning),} \\ r_{\Theta, B}(\boldsymbol{\Theta}, \mathbf{B}) = \delta(\boldsymbol{\Theta}; \widehat{\boldsymbol{\Theta}}) r_B(\mathbf{B}) & \text{(for PB-B learning),} \\ r_{\Theta, B}(\boldsymbol{\Theta}, \mathbf{B}) = r_{\Theta}(\boldsymbol{\Theta}) r_B(\mathbf{B}) & \text{(for VB learning),} \end{cases}$$

Similar analysis to Section 16.4.1 leads to the following theorem (the proof is given in Section 16.4.5):

Theorem 16.13 *In the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$, the solution is sparse if $\alpha < \underline{\alpha}_{\text{sparse}}$, and dense if $\alpha > \underline{\alpha}_{\text{dense}}$. In the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$, the solution is sparse if $\alpha < \underline{\alpha}_{M \rightarrow \infty}$, and dense if $\alpha > \underline{\alpha}_{M \rightarrow \infty}$. Here, $\underline{\alpha}_{\text{sparse}}$, $\underline{\alpha}_{\text{dense}}$, and $\underline{\alpha}_{M \rightarrow \infty}$ are given in Table 16.1.*

A notable finding from Table 16.1 is that the threshold that determines the topic sparsity of PB-B learning is (most of the case exactly) $\frac{1}{2}$ larger than the threshold of VB learning. The same relation is observed between MAP learning and PB-A learning. From these, we can conclude that point-estimating $\boldsymbol{\Theta}$, instead of integrating it out, increases the threshold by $\frac{1}{2}$ in the LDA model. We will validate this observation by numerical experiments in Section 16.4.4.

Table 16.1 *Sparsity thresholds of VB, PB-A, PB-B, and MAP methods (see Theorem 16.13). The first four columns show the thresholds $(\underline{\alpha}_{\text{sparse}}, \underline{\alpha}_{\text{dense}})$, of which the function forms depend on the range of η , in the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$. A single value is shown if $\underline{\alpha}_{\text{sparse}} = \underline{\alpha}_{\text{dense}}$. The last column shows the threshold $\underline{\alpha}_{M \rightarrow \infty}$ in the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$.*

η range	$(\underline{\alpha}_{\text{sparse}}, \underline{\alpha}_{\text{dense}})$				$\underline{\alpha}_{M \rightarrow \infty}$
	$0 < \eta \leq \frac{1}{2L}$	$\frac{1}{2L} < \eta \leq \frac{1}{2}$	$\frac{1}{2} < \eta < 1$	$1 \leq \eta < \infty$	
VB	$\frac{1}{2} - \frac{\frac{1}{2} - L\eta}{\min_h M^{*(h)}}$	$\frac{1}{2} + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}$	$\left(\frac{1}{2} + \frac{L-1}{2 \max_h M^{*(h)}}, \frac{1}{2} + \frac{L\eta - \frac{1}{2}}{\min_h M^{*(h)}} \right)$		$\frac{1}{2}$
PB-A		—		$\left(\frac{1}{2}, \frac{1}{2} + \frac{L(\eta-1)}{\min_h M^{*(h)}} \right)$	$\frac{1}{2}$
PB-B	1	$1 + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}$	$\left(1 + \frac{L-1}{2 \max_h M^{*(h)}}, 1 + \frac{L\eta - \frac{1}{2}}{\min_h M^{*(h)}} \right)$		1
MAP		—		$\left(1, 1 + \frac{L(\eta-1)}{\min_h M^{*(h)}} \right)$	1

16.4.3 Discussion

The preceding theoretical analysis (Thereom 16.13) showed that VB tends to induce weaker sparsity than MAP in the LDA model,³ i.e., VB requires sparser prior (smaller α) than MAP to give a sparse solution (mean of the posterior). This phenomenon is opposite to other models such as mixture models (Chapter 15), Bayesian networks (Section 16.1), hidden Markov models (Section 16.2), and fully observed matrix factorization (Chapter 7), where VB tends to induce stronger sparsity than MAP. This phenomenon might be partly explained as follows: in the case of mixture models, the sparsity threshold depends on the degree of freedom of a single component (Theorem 15.5). This is reasonable because adding a single component increases the model complexity by this amount. Also, in the case of LDA, adding a single topic requires additional $L+1$ parameters. However, the added topic is shared over M documents, which could discount the increased model complexity relative to the increased data fidelity. Corollary 16.11, which implies the dependency of the threshold for α on L and M , might support this conjecture. However, the same applies to the matrix factorization, where VB was shown to give a sparser solution than MAP (Chapter 7). Investigation on related models, e.g., Poisson MF (Gopalan et al., 2013), would help us fully explain this phenomenon.

Unlike for the latent variable models in the previous sections, we derived a general form of the asymptotic free energy for LDA, which can be applied to different asymptotic limits and showed that the consistency does not always hold (see Theorem 16.10). Specifically, the standard asymptotic theory requires a large number N of words per document, compared to the number M of documents and the vocabulary size L . Assuming such a situation may be reasonable in some collaborative filtering applications, e.g., in the *Last.FM* data which will be used for numerical illustration in Section 16.4.4. However, L and/or M are comparable to or larger than N in many text analysis applications.

The general form of the asymptotic free energy also allowed us to elucidate the behavior of the VB free energy when L and/or M diverges with the same order as N . This attempt successfully revealed the sparsity of the solution for the case when M diverges while $L \sim O(1)$. However, when L diverges, we found that the leading term of the free energy does not contain useful information on sparsity of the solution. Higher-order asymptotic analysis will be necessary to further understand the sparsity-inducing mechanism of the LDA model with large vocabulary.

³ This tendency was pointed out (Asuncion et al., 2009) by using the approximation $\exp(\Psi(n)) \approx n - \frac{1}{2}$ and comparing the stationary condition. The theory here clarified the sparsity behavior of the global solution based on the asymptotic free energy analysis.

16.4.4 Numerical Illustration

Here we conduct numerical experiments on artificial and real data for collaborative filtering.

The *artificial* data were created as follows: we first sample the *true* document matrix Θ^* of size $M \times H^*$ and the *true* topic matrix B^* of size $L \times H^*$. We assume that each row $\tilde{\theta}_m^*$ of Θ^* follows the Dirichlet distribution with $\alpha^* = 1/H^*$, while each column β_h^* of B^* follows the Dirichlet distribution with $\eta^* = 1/L$. The document length $N^{(m)}$ is sampled from the Poisson distribution with mean N . The word histogram $N^{(m)}v_m$ for each document is sampled from the multinomial distribution with the parameter specified by the m th row vector of $B^*\Theta^{*\top}$. Thus, we obtain the $L \times M$ matrix V , which corresponds to the empirical word distribution over M documents.

As a real-world data set, we used the *Last.FM* data set.⁴ *Last.FM* is a well-known social music web site, and the data set includes the triple (“user,” “artist,” “Freq”), which was collected from the playlists of users in

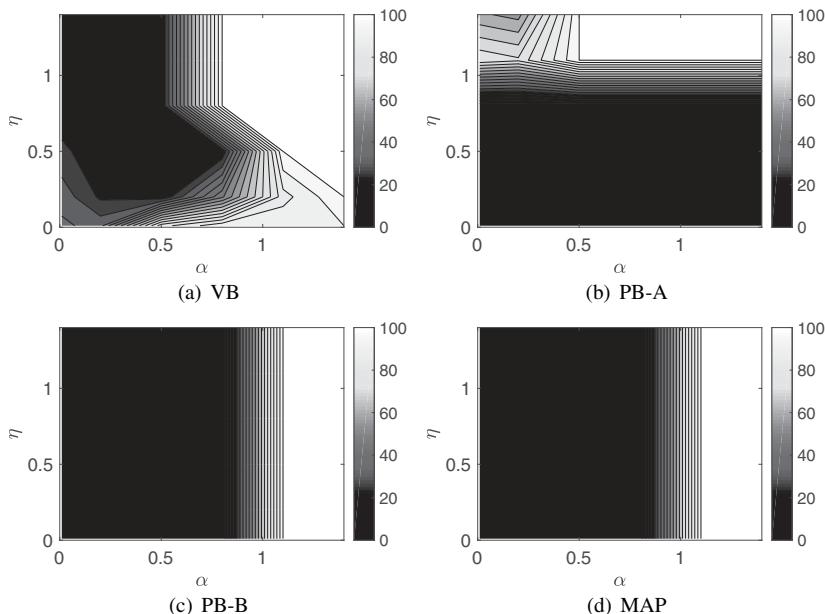


Figure 16.2 Estimated number \widehat{H} of topics by (a) VB learning, (b) PB-A learning, (c) PB-B learning, and (d) MAP learning, on the *artificial* data with $L = 100$, $M = 100$, $H^* = 20$, and $N \sim 10000$.

⁴ <http://mtg.upf.edu/node/1671>

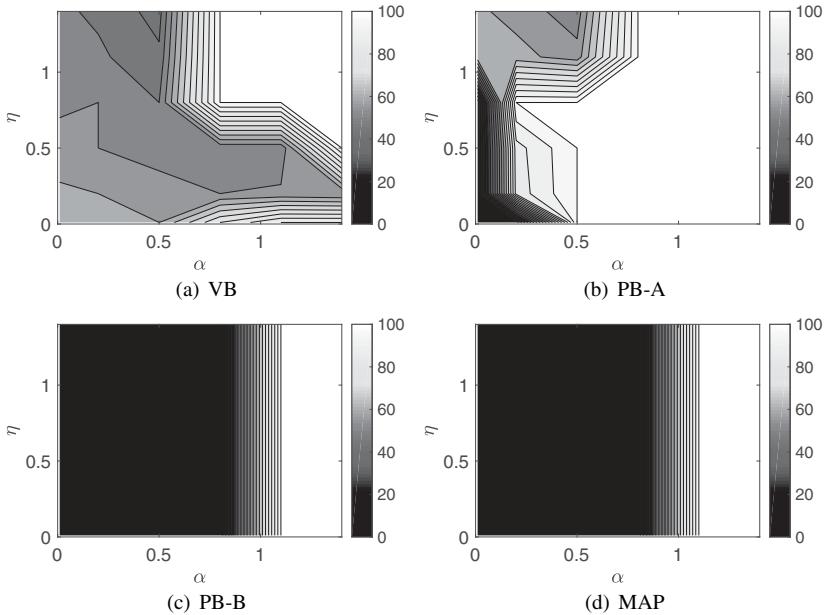


Figure 16.3 Estimated number \widehat{H} of topics on the *Last.FM* data with $L = 100$, $M = 100$, and $N \sim 700$.

the community by using a plug-in in users' media players. This triple means that "user" played "artist" music "Freq" times, which indicates users' preferred artists. A user and a played artist are analogous to a document and a word, respectively. We randomly chose L artists from the top 1,000 frequent artists, and M users who live in the United States. To find a better local solution (which hopefully is close to the global solution), we adopted a split and merge strategy (Ueda et al., 2000), and chose the local solution giving the lowest free energy among different initialization schemes.

Figure 16.2 shows the estimated number \widehat{H} of topics by different approximate Bayesian methods, i.e., VB, PB-A, PB-B, and MAP learning, on the *artificial* data with $L = 100$, $M = 100$, $H^* = 20$, and $N \sim 10000$. We can clearly see that the sparsity threshold in PB-B and MAP learning, where Θ is point-estimated, is larger than that in VB and PB-A learning, where Θ is marginalized. This result supports the statement by Theorem 16.13. Figure 16.3 shows results on the *Last.FM* data with $L = 100$, $M = 100$, and $N \sim 700$. We see a similar tendency to Figure 16.2 except the region where $\eta < 1$ for PB-A learning, in which our theory does not predict the estimated number of topics.

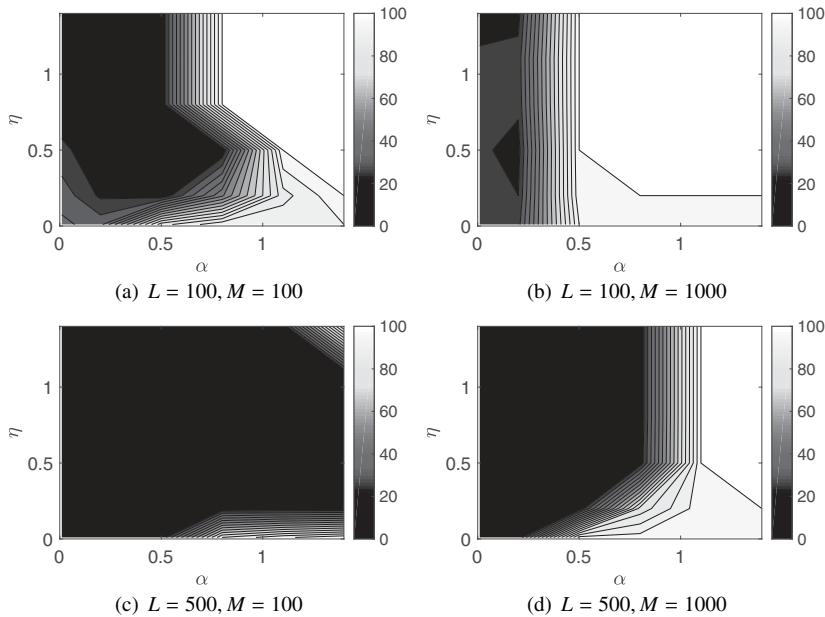


Figure 16.4 Estimated number \widehat{H} of topics by VB learning on the *artificial* data with $H^* = 20$ and $N \sim 10000$. For the case when $L = 500, M = 1000$, the maximum estimated rank is limited to 100 for computational reason.

Finally, we investigate how different asymptotic settings affect the topic sparsity. Figure 16.4 shows the sparsity dependence on L and M on the *artificial* data. The graphs correspond to the four cases mentioned in Theorem 16.10, i.e., (a) $L, M \ll N$, (b) $L \ll N, M$, (c) $M \ll N, L$, and (d) $1 \ll N, L, M$. Corollary 16.11 explains the behavior in (a) and (b), and further analysis is required to explain the behavior in (c) and (d).

16.4.5 Proof of Theorem 16.13

We analyze PB-A learning, PB-B learning, and MAP learning, and then summarize the results, which proves Theorem 16.13.

PB-A Learning

The free energy for PB-A learning is given as follows:

$$F^{\text{PB-A}} = \chi_B + R^{\text{PB-A}} + Q^{\text{PB-A}}, \quad (16.64)$$

where χ_B is a large constant corresponding to the negative entropy of the delta functions (see Section 2.2.2), and

$$\begin{aligned} R^{\text{PB-A}} &= \left\langle \log \frac{r_{\Theta}(\boldsymbol{\Theta}) r_B(\mathbf{B})}{p(\boldsymbol{\Theta}|\alpha) p(\mathbf{B}|\eta)} \right\rangle_{r^{\text{PB-A}}(\boldsymbol{\Theta}, \mathbf{B})} \\ &= \sum_{m=1}^M \left(\log \frac{\Gamma(\sum_{h=1}^H \widehat{\alpha}_{m,h}^{\text{PB-A}})}{\prod_{h=1}^H \Gamma(\widehat{\alpha}_{m,h}^{\text{PB-A}})} \frac{\Gamma(\alpha)^H}{\Gamma(H\alpha)} \right. \\ &\quad \left. + \sum_{h=1}^H (\widehat{\alpha}_{m,h}^{\text{PB-A}} - \alpha) \left(\Psi(\widehat{\alpha}_{m,h}^{\text{PB-A}}) - \Psi(\sum_{h'=1}^H \widehat{\alpha}_{m,h'}^{\text{PB-A}}) \right) \right) \\ &\quad + \sum_{h=1}^H \left(\log \frac{\Gamma(\eta)^L}{\Gamma(L\eta)} + \sum_{l=1}^L (1-\eta) (\log(\widehat{\eta}_{l,h}^{\text{PB-A}}) - \log(\sum_{l'=1}^L \widehat{\eta}_{l',h}^{\text{PB-A}})) \right), \end{aligned} \quad (16.65)$$

$$\begin{aligned} Q^{\text{PB-A}} &= \left\langle \log \frac{r_z\left(\{\{\mathbf{z}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M\right)}{p(\{\mathbf{w}^{(n,m)}\}, \{\mathbf{z}^{(n,m)}\} | \boldsymbol{\Theta}, \mathbf{B})} \right\rangle_{r^{\text{PB-A}}(\boldsymbol{\Theta}, \mathbf{B}, \{\mathbf{z}^{(n,m)}\})} \\ &= - \sum_{m=1}^M N^{(m)} \sum_{l=1}^L V_{l,m} \log \left(\sum_{h=1}^H \frac{\exp(\Psi(\widehat{\alpha}_{m,h}^{\text{PB-A}}))}{\exp(\Psi(\sum_{h'=1}^H \widehat{\alpha}_{m,h'}^{\text{PB-A}}))} \frac{\widehat{\eta}_{l,h}^{\text{PB-A}}}{\sum_{l'=1}^L \widehat{\eta}_{l',h}^{\text{PB-A}}} \right). \end{aligned} \quad (16.66)$$

Let us first consider the case when $\eta < 1$. In this case, F diverges to $F \rightarrow -\infty$ with fixed N , when $\widehat{\eta}_{l,h} = O(1)$ for any (l, h) and $\widehat{\eta}_{l',h} \rightarrow +0$ for all other $l' \neq l$. Therefore, the solution is useless.

When $\eta \geq 1$, the solution satisfies the following stationary condition:

$$\widehat{\alpha}_{m,h}^{\text{PB-A}} = \alpha + \sum_{n=1}^{N^{(m)}} \widehat{z}_h^{\text{PB-A}(n,m)}, \quad \widehat{\eta}_{l,h}^{\text{PB-A}} = \eta - 1 + \sum_{m=1}^M \sum_{n=1}^{N^{(m)}} w_l^{(n,m)} \widehat{z}_h^{\text{PB-A}(n,m)}, \quad (16.67)$$

$$\widehat{z}_h^{\text{PB-A}(n,m)} = \frac{\exp(\Psi(\widehat{\alpha}_{m,h}^{\text{PB-A}})) \prod_{l=1}^L (\widehat{\eta}_{l,h}^{\text{PB-A}})^{w_l^{(n,m)}}}{\sum_{h'=1}^H \left(\exp(\Psi(\widehat{\alpha}_{m,h'}^{\text{PB-A}})) \prod_{l=1}^L (\widehat{\eta}_{l,h'}^{\text{PB-A}})^{w_l^{(n,m)}} \right)}. \quad (16.68)$$

In the same way as for VB learning, we can obtain the following lemma:

Lemma 16.14 *Let $\widehat{\mathbf{B}}^{\text{PB-A}} \widehat{\boldsymbol{\theta}}^{\text{PB-A}\top} = \langle \mathbf{B} \boldsymbol{\theta}^\top \rangle_{r^{\text{PB-A}}(\boldsymbol{\Theta}, \mathbf{B})}$. Then it holds that*

$$\langle (\mathbf{B} \boldsymbol{\theta}^\top - \widehat{\mathbf{B}}^{\text{PB-A}} \widehat{\boldsymbol{\theta}}^{\text{PB-A}\top})_{l,m}^2 \rangle_{r^{\text{PB-A}}(\boldsymbol{\Theta}, \mathbf{B})} = O_p(N^{-2}), \quad (16.69)$$

$$Q^{\text{PB-A}} = - \sum_{m=1}^M N^{(m)} \sum_{l=1}^L V_{l,m} \log \langle \widehat{\mathbf{B}}^{\text{PB-A}} \widehat{\boldsymbol{\theta}}^{\text{PB-A}\top} \rangle_{l,m} + O_p(N^{-1}). \quad (16.70)$$

$Q^{\text{PB-A}}$ is minimized when $\widehat{J} = O_p(N^{-1})$, and it holds that

$$Q^{\text{PB-A}} = NS_N(\mathcal{D}) + O_p(\widehat{J}N + LM).$$

$R^{\text{PB-A}}$ is written as follows:

$$\begin{aligned} R^{\text{PB-A}} &= \left\{ M \left(H\alpha - \frac{1}{2} \right) + \widehat{H}L(\eta - 1) - \sum_{h=1}^{\widehat{H}} \left(\widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) + \widehat{L}^{(h)}(\eta - 1) \right) \right\} \log N \\ &\quad + (H - \widehat{H})L(\eta - 1) \log L + O_p(H(M + L)). \end{aligned} \quad (16.71)$$

Taking the different asymptotic limits, we obtain the following theorem:

Theorem 16.15 When $\eta < 1$, each column vector of $\widehat{\mathbf{B}}^{\text{PB-A}}$ has only one nonzero entry. Assume in the following that $\eta \geq 1$. In the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$, it holds that $\widehat{J} = O_p(1/N)$ and

$$\widetilde{F}^{\text{PB-A}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{PB-A}} \log N + O_p(1),$$

where

$$\lambda'_{\text{LDA}}^{\text{PB-A}} = M \left(H\alpha - \frac{1}{2} \right) + \widehat{H}L(\eta - 1) - \sum_{h=1}^{\widehat{H}} \left(\widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) + \widehat{L}^{(h)}(\eta - 1) \right).$$

In the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$, it holds that $\widehat{J} = o_p(\log N)$, and

$$\widetilde{F}^{\text{PB-A}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{PB-A}} \log N + o_p(N \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{PB-A}} = M \left(H\alpha - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right).$$

In the limit when $N, L \rightarrow \infty$ with $\frac{L}{N}, M \sim O(1)$, it holds that $\widehat{J} = o_p(\log N)$, and

$$\widetilde{F}^{\text{PB-A}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{PB-A}} \log N + o_p(N \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{PB-A}} = HL(\eta - 1).$$

In the limit when $N, L, M \rightarrow \infty$ with $\frac{L}{N}, \frac{M}{N} \sim O(1)$, it holds that $\widehat{J} = o_p(N \log N)$, and

$$\widetilde{F}^{\text{PB-A}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{PB-A}} \log N + o_p(N^2 \log N),$$

where

$$\lambda_{\text{LDA}}^{\text{PB-A}} = H(M\alpha + L(\eta - 1)).$$

Note that Theorem 16.15 provides no information on the sparsity of the PB-A solution for $\eta < 1$. In the following subsection, we investigate the sparsity of the solution for $\eta \geq 1$.

In the Limit When $N \rightarrow \infty$ with $L, M \sim O(1)$

The coefficient of the leading term of the free energy is

$$\lambda_{\text{LDA}}^{\text{PB-A}} = M \left(H\alpha - \frac{1}{2} \right) + \sum_{h=1}^{\widehat{H}} \left(L(\eta - 1) - \widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right) - \widehat{L}^{(h)} (\eta - 1) \right).$$

The solution is sparse if $\lambda_{\text{LDA}}^{\text{PB-A}}$ is increasing with respect to \widehat{H} , and dense if it is decreasing. We focus on the case where $\eta \geq 1$. Eqs. (16.57) through (16.59) imply the following:

(I) When $\alpha < \frac{1}{2}$, the solution is sparse if

$$L(\eta - 1) - \max_h \left(M^{*(h)} \left(\alpha - \frac{1}{2} \right) + L^{*(h)} (\eta - 1) \right) > 0, \quad (16.72)$$

and dense if

$$L(\eta - 1) - \max_h \left(M^{*(h)} \left(\alpha - \frac{1}{2} \right) + L^{*(h)} (\eta - 1) \right) < 0. \quad (16.73)$$

Therefore, the solution is sparse if

$$L(\eta - 1) - \min_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) - \max_h L^{*(h)} (\eta - 1) > 0, \text{ or equivalently,}$$

$$\alpha < \frac{1}{2} + \frac{(L - \max_h L^{*(h)}) (\eta - 1)}{\min_h M^{*(h)}},$$

and dense if

$$L(\eta - 1) - \max_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) - \max_h L^{*(h)} (\eta - 1) < 0, \text{ or equivalently,}$$

$$\alpha > \frac{1}{2} + \frac{(L - \max_h L^{*(h)}) (\eta - 1)}{\max_h M^{*(h)}}.$$

Therefore, the solution is always sparse in this case.

- (II) When $\alpha \geq \frac{1}{2}$, the solution is sparse if Eq. (16.72) holds, and dense if Eq. (16.73) holds. Therefore, the solution is sparse if

$$L(\eta - 1) - \max_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) - \max_h L^{*(h)} (\eta - 1) > 0, \text{ or equivalently,}$$

$$\alpha < \frac{1}{2} + \frac{(L - \max_h L^{*(h)}) (\eta - 1)}{\max_h M^{*(h)}},$$

and dense if

$$L(\eta - 1) - \min_h M^{*(h)} \left(\alpha - \frac{1}{2} \right) - \max_h L^{*(h)} (\eta - 1) < 0, \text{ or equivalently,}$$

$$\alpha > \frac{1}{2} + \frac{(L - \max_h L^{*(h)}) (\eta - 1)}{\min_h M^{*(h)}}.$$

Thus, we can conclude that, in this case, the solution is sparse if

$$\alpha < \frac{1}{2},$$

and dense if

$$\alpha > \frac{1}{2} + \frac{L(\eta - 1)}{\min_h M^{*(h)}}.$$

Summarizing the preceding, we have the following lemma:

Lemma 16.16 *Assume that $\eta \geq 1$. The solution is sparse if $\alpha < \frac{1}{2}$, and dense if $\alpha > \frac{1}{2} + \frac{L(\eta - 1)}{\min_h M^{*(h)}}$.*

In the Limit When $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$

The coefficient of the leading term of the free energy is given by

$$\lambda_{\text{LDA}}^{\text{PB-A}} = M \left(H\alpha - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \widehat{M}^{(h)} \left(\alpha - \frac{1}{2} \right). \quad (16.74)$$

Although the predictive distribution does not necessarily converge to the true distribution, we can investigate the sparsity of the solution by considering the duplication rules (16.57) through (16.59) that keep $\widehat{\mathbf{B}}\widehat{\boldsymbol{\Theta}}^\top$ unchanged. It is clear that Eq. (16.74) is increasing with respect to \widehat{H} if $\alpha < \frac{1}{2}$, and decreasing if $\alpha > \frac{1}{2}$. Combing this result with Lemma 16.16, we obtain the following corollary:

Corollary 16.17 *Assume that $\eta \geq 1$. In the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$, the PB-A solution is sparse if $\alpha < \frac{1}{2}$, and dense if $\alpha > \frac{1}{2} + \frac{L(\eta - 1)}{\min_h M^{*(h)}}$.*

In the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$, the PB-A solution is sparse if $\alpha < \frac{1}{2}$, and dense if $\alpha > \frac{1}{2}$.

PB-B Learning

The free energy for PB-B learning is given as follows:

$$F^{\text{PB-B}} = \chi_\theta + R^{\text{PB-B}} + Q^{\text{PB-B}}, \quad (16.75)$$

where χ_θ is a large constant corresponding to the negative entropy of the delta functions, and

$$\begin{aligned} R^{\text{PB-B}} &= \left\langle \log \frac{r_\Theta(\boldsymbol{\Theta}) r_B(\mathbf{B})}{p(\boldsymbol{\Theta}|\alpha) p(\mathbf{B}|\eta)} \right\rangle_{r^{\text{PB-B}}(\boldsymbol{\Theta}, \mathbf{B})} \\ &= \sum_{m=1}^M \left(\log \frac{\Gamma(\alpha)^H}{\Gamma(H\alpha)} + \sum_{h=1}^H (1-\alpha) \left(\log(\widehat{\alpha}_{m,h}^{\text{PB-B}}) - \log \left(\sum_{h'=1}^H \widehat{\alpha}_{m,h'}^{\text{PB-B}} \right) \right) \right) \\ &\quad + \sum_{h=1}^H \left(\log \frac{\Gamma(\sum_{l=1}^L \widehat{\eta}_{l,h}^{\text{PB-B}})}{\prod_{l=1}^L \Gamma(\widehat{\eta}_{l,h}^{\text{PB-B}})} \frac{\Gamma(\eta)^L}{\Gamma(L\eta)} \right. \\ &\quad \left. + \sum_{l=1}^L (\widehat{\eta}_{l,h}^{\text{PB-B}} - \eta) (\Psi(\widehat{\eta}_{l,h}^{\text{PB-B}}) - \Psi(\sum_{l'=1}^L \widehat{\eta}_{l',h}^{\text{PB-B}})) \right), \end{aligned} \quad (16.76)$$

$$\begin{aligned} Q^{\text{PB-B}} &= \left\langle \log \frac{r_z \left(\{\{\mathbf{z}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M \right)}{p(\{\mathbf{w}^{(n,m)}\}, \{\mathbf{z}^{(n,m)}\} | \boldsymbol{\Theta}, \mathbf{B})} \right\rangle_{r^{\text{PB-B}}(\boldsymbol{\Theta}, \mathbf{B}, \{\mathbf{z}^{(n,m)}\})} \\ &= - \sum_{m=1}^M N^{(m)} \sum_{l=1}^L V_{l,m} \log \left(\sum_{h=1}^H \frac{\widehat{\alpha}_{m,h}^{\text{PB-B}}}{\sum_{h'=1}^H \widehat{\alpha}_{m,h'}^{\text{PB-B}}} \frac{\exp(\Psi(\widehat{\eta}_{l,h}^{\text{PB-B}}))}{\exp(\Psi(\sum_{l'=1}^L \widehat{\eta}_{l',h}^{\text{PB-B}}))} \right). \end{aligned} \quad (16.77)$$

Let us first consider the case when $\alpha < 1$. In this case, F diverges to $F \rightarrow -\infty$ with fixed N , when $\widehat{\alpha}_{m,h} = O(1)$ for any (m, h) and $\widehat{\alpha}_{m,h'} \rightarrow +0$ for all other $h' \neq h$. Therefore, the solution is sparse (so sparse that the estimator is useless).

When $\alpha \geq 1$, the solution satisfies the following stationary condition:

$$\widehat{\alpha}_{m,h}^{\text{PB-B}} = \alpha - 1 + \sum_{n=1}^{N^{(m)}} \widehat{z}_h^{\text{PB-B}(n,m)}, \quad \widehat{\eta}_{l,h}^{\text{PB-B}} = \eta + \sum_{m=1}^M \sum_{n=1}^{N^{(m)}} w_l^{(n,m)} \widehat{z}_h^{\text{PB-B}(n,m)}, \quad (16.78)$$

$$\widehat{z}_h^{\text{PB-B}(n,m)} = \frac{\widehat{\alpha}_{m,h}^{\text{PB-B}} \exp \left\{ \sum_{l=1}^L w_l^{(n,m)} (\Psi(\widehat{\eta}_{l,h}^{\text{PB-B}}) - \Psi(\sum_{l'=1}^L \widehat{\eta}_{l',h}^{\text{PB-B}})) \right\}}{\sum_{h'=1}^H \widehat{\alpha}_{m,h'}^{\text{PB-B}} \exp \left\{ \sum_{l=1}^L w_l^{(n,m)} (\Psi(\widehat{\eta}_{l,h'}^{\text{PB-B}}) - \Psi(\sum_{l'=1}^L \widehat{\eta}_{l',h'}^{\text{PB-B}})) \right\}}. \quad (16.79)$$

In the same way as for VB and PB-A learning, we can obtain the following lemma:

Lemma 16.18 *Let $\widehat{\mathbf{B}}^{\text{PB-B}} \widehat{\boldsymbol{\theta}}^{\text{PB-B}\top} = \langle \mathbf{B} \boldsymbol{\theta}^\top \rangle_{r^{\text{PB-B}}(\boldsymbol{\theta}, \mathbf{B})}$. Then it holds that*

$$\langle (\mathbf{B} \boldsymbol{\theta}^\top - \widehat{\mathbf{B}}^{\text{PB-B}} \widehat{\boldsymbol{\theta}}^{\text{PB-B}\top})_{l,m}^2 \rangle_{r^{\text{PB-B}}(\boldsymbol{\theta}, \mathbf{B})} = O_p(N^{-2}), \quad (16.80)$$

$$Q^{\text{PB-B}} = - \sum_{m=1}^M N^{(m)} \sum_{l=1}^L V_{l,m} \log(\widehat{\mathbf{B}}^{\text{PB-B}} \widehat{\boldsymbol{\theta}}^{\text{PB-B}\top})_{l,m} + O_p(N^{-1}). \quad (16.81)$$

$Q^{\text{PB-B}}$ is minimized when $\widehat{J} = O_p(N^{-1})$, and it holds that

$$Q^{\text{PB-B}} = NS_N(\mathcal{D}) + O_p(\widehat{J}N + LM).$$

$R^{\text{PB-B}}$ is written as follows:

$$R^{\text{PB-B}} = \left\{ MH(\alpha - 1) + \widehat{H} \left(L\eta - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \left(\widehat{M}^{(h)}(\alpha - 1) + \widehat{L}^{(h)} \left(\eta - \frac{1}{2} \right) \right) \right\} \log N + (H - \widehat{H}) \left(L\eta - \frac{1}{2} \right) \log L + O_p(H(M + L)). \quad (16.82)$$

Taking the different asymptotic limits, we obtain the following theorem:

Theorem 16.19 *When $\alpha < 1$, each row vector of $\widehat{\boldsymbol{\theta}}^{\text{PB-B}}$ has only one nonzero entry, and the PB-B solution is sparse. Assume in the following that $\alpha \geq 1$. In the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$, it holds that $\widehat{J} = O_p(1/N)$ and*

$$\widetilde{F}^{\text{PB-B}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{PB-B}} \log N + O_p(1),$$

where

$$\lambda'_{\text{LDA}}^{\text{PB-B}} = MH(\alpha - 1) + \widehat{H} \left(L\eta - \frac{1}{2} \right) - \sum_{h=1}^{\widehat{H}} \left(\widehat{M}^{(h)}(\alpha - 1) + \widehat{L}^{(h)} \left(\eta - \frac{1}{2} \right) \right).$$

In the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$, it holds that $\widehat{J} = o_p(\log N)$, and

$$\widetilde{F}^{\text{PB-B}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{PB-B}} \log N + o_p(N \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{PB-B}} = MH(\alpha - 1) - \sum_{h=1}^{\widehat{H}} \widehat{M}^{(h)}(\alpha - 1).$$

In the limit when $N, L \rightarrow \infty$ with $\frac{L}{N}, M \sim O(1)$, it holds that $\widehat{J} = o_p(\log N)$, and

$$\widetilde{F}^{\text{PB-B}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{PB-B}} \log N + o_p(N \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{PB-B}} = HL\eta.$$

In the limit when $N, L, M \rightarrow \infty$ with $\frac{L}{N}, \frac{M}{N} \sim O(1)$, it holds that $\widehat{J} = o_p(N \log N)$, and

$$\widetilde{F}^{\text{PB-B}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{PB-B}} \log N + o_p(N^2 \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{PB-B}} = H(M(\alpha - 1) + L\eta).$$

Theorem 16.19 states that the PB-B solution is sparse when $\alpha < 1$. In the following subsection, we investigate the sparsity of the solution for $\alpha \geq 1$.

In the Limit When $N \rightarrow \infty$ with $L, M \sim O(1)$

The coefficient of the leading term of the free energy is

$$\lambda'_{\text{LDA}}^{\text{PB-B}} = MH(\alpha - 1) + \sum_{h=1}^{\widehat{H}} \left(L\eta - \frac{1}{2} - \widehat{M}^{(h)}(\alpha - 1) - \widehat{L}^{(h)} \left(\eta - \frac{1}{2} \right) \right).$$

The solution is sparse if $\lambda'_{\text{LDA}}^{\text{PB-B}}$ is increasing with respect to \widehat{H} , and dense if it is decreasing. We focus on the case where $\alpha \geq 1$. Eqs. (16.57) through (16.59) imply the following:

(I) When $0 < \eta \leq \frac{1}{2L}$, the solution is sparse if

$$\begin{aligned} L\eta - \frac{1}{2} - \max_h M^{*(h)}(\alpha - 1) &> 0, \text{ or equivalently,} \\ \alpha &< 1 - \frac{1}{\max_h M^{*(h)}} \left(\frac{1}{2} - L\eta \right), \end{aligned}$$

and dense if

$$\alpha > 1 - \frac{1}{\max_h M^{*(h)}} \left(\frac{1}{2} - L\eta \right).$$

Therefore, the solution is always dense in this case.

(II) When $\frac{1}{2L} < \eta \leq \frac{1}{2}$, the solution is sparse if

$$L\eta - \frac{1}{2} - \max_h M^{*(h)}(\alpha - 1) > 0, \text{ or equivalently, } \alpha < 1 + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}},$$

and dense if

$$\alpha > 1 + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}.$$

(III) When $\eta > \frac{1}{2}$, the solution is sparse if

$$L\eta - \frac{1}{2} - \max_h \left(M^{*(h)}(\alpha - 1) + L^{*(h)}\left(\eta - \frac{1}{2}\right) \right) > 0, \quad (16.83)$$

and dense if

$$L\eta - \frac{1}{2} - \max_h \left(M^{*(h)}(\alpha - 1) + L^{*(h)}\left(\eta - \frac{1}{2}\right) \right) < 0. \quad (16.84)$$

Therefore, the solution is sparse if

$$\begin{aligned} L\eta - \frac{1}{2} - \max_h M^{*(h)}(\alpha - 1) - \max_h L^{*(h)}\left(\eta - \frac{1}{2}\right) &> 0, \text{ or equivalently,} \\ \alpha &< 1 + \frac{1}{\max_h M^{*(h)}} \left(L\eta - \frac{1}{2} - \max_h L^{*(h)}\left(\eta - \frac{1}{2}\right) \right), \end{aligned}$$

and dense if

$$\begin{aligned} L\eta - \frac{1}{2} - \min_h M^{*(h)}(\alpha - 1) - \max_h L^{*(h)}\left(\eta - \frac{1}{2}\right) &< 0, \text{ or equivalently,} \\ \alpha &> 1 + \frac{1}{\min_h M^{*(h)}} \left(L\eta - \frac{1}{2} - \max_h L^{*(h)}\left(\eta - \frac{1}{2}\right) \right). \end{aligned}$$

Thus, we can conclude that, in this case, the solution is sparse if

$$\alpha < 1 + \frac{L - 1}{2 \max_h M^{*(h)}},$$

and dense if

$$\alpha > 1 + \frac{L\eta - \frac{1}{2}}{\min_h M^{*(h)}}.$$

Summarizing the preceding, we have the following lemma:

Lemma 16.20 Assume that $\alpha \geq 1$. When $0 < \eta \leq \frac{1}{2L}$, the solution is always dense. When $\frac{1}{2L} < \eta \leq \frac{1}{2}$, the solution is sparse if $\alpha < 1 + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}$, and dense if $\alpha > 1 + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}$. When $\eta > \frac{1}{2}$, the solution is sparse if $\alpha < 1 + \frac{L - 1}{2 \max_h M^{*(h)}}$, and dense if $\alpha > 1 + \frac{L\eta - \frac{1}{2}}{\min_h M^{*(h)}}$.

In the Limit When $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$

The coefficient of the leading term of the free energy is given by

$$\lambda'_{\text{LDA}}^{\text{PB-B}} = M(H\alpha - 1) - \sum_{h=1}^{\widehat{H}} \widehat{M}^{(h)}(\alpha - 1). \quad (16.85)$$

Although the predictive distribution does not necessarily converge to the true distribution, we can investigate the sparsity of the solution by considering the duplication rules (16.57) through (16.59) that keep $\widehat{\mathbf{B}\Theta}^\top$ unchanged. It is clear that Eq. (16.85) is decreasing with respect to \widehat{H} if $\alpha > 1$. Combing this result with Theorem 16.19, which states that the PB-B solution is sparse when $\alpha < 1$, and Lemma 16.20, we obtain the following corollary:

Corollary 16.21 *Consider the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$. When $0 < \eta \leq \frac{1}{2L}$, the PB-B solution is sparse if $\alpha < 1$, and dense if $\alpha > 1$. When $\frac{1}{2L} < \eta \leq \frac{1}{2}$, the PB-B solution is sparse if $\alpha < 1 + \frac{L\eta - \frac{1}{2}}{\max_h M^{*(h)}}$, and dense if $\alpha > 1 + \frac{L\eta - \frac{1}{2}}{\min_h M^{*(h)}}$. When $\eta > \frac{1}{2}$, the PB-B solution is sparse if $\alpha < 1 + \frac{L-1}{2\max_h M^{*(h)}}$, and dense if $\alpha > 1 + \frac{L\eta - \frac{1}{2}}{\min_h M^{*(h)}}$. In the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$, the PB-B solution is sparse if $\alpha < 1$, and dense if $\alpha > 1$.*

MAP Learning

The free energy for MAP learning is given as follows:

$$F^{\text{MAP}} = \chi_\Theta + \chi_B + R^{\text{MAP}} + Q^{\text{MAP}}, \quad (16.86)$$

where χ_Θ and χ_B are large constants corresponding to the negative entropies of the delta functions, and

$$\begin{aligned} R^{\text{MAP}} &= \left\langle \log \frac{r_\Theta(\boldsymbol{\Theta}) r_B(\mathbf{B})}{p(\boldsymbol{\Theta}|\alpha)p(\mathbf{B}|\eta)} \right\rangle_{r^{\text{MAP}}(\boldsymbol{\Theta}, \mathbf{B})} \\ &= \sum_{m=1}^M \left(\log \frac{\Gamma(\alpha)^H}{\Gamma(H\alpha)} + \sum_{h=1}^H (1-\alpha) \left(\log(\widehat{\alpha}_{m,h}^{\text{MAP}}) - \log(\sum_{h'=1}^H \widehat{\alpha}_{m,h'}^{\text{MAP}}) \right) \right) \\ &\quad + \sum_{h=1}^H \left(\log \frac{\Gamma(\eta)^L}{\Gamma(L\eta)} + \sum_{l=1}^L (1-\eta) \left(\log(\widehat{\eta}_{l,h}^{\text{MAP}}) - \log(\sum_{l'=1}^L \widehat{\eta}_{l',h}^{\text{MAP}}) \right) \right), \end{aligned} \quad (16.87)$$

$$\begin{aligned} Q^{\text{MAP}} &= \left\langle \log \frac{r_z(\{\{z^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^M)}{p(\{\mathbf{w}^{(n,m)}\}, \{z^{(n,m)}\} | \boldsymbol{\Theta}, \mathbf{B})} \right\rangle_{r^{\text{MAP}}(\boldsymbol{\Theta}, \mathbf{B}, \{z^{(n,m)}\})} \\ &= - \sum_{m=1}^M N^{(m)} \sum_{l=1}^L V_{l,m} \log \left(\sum_{h=1}^H \frac{\widehat{\alpha}_{m,h}^{\text{MAP}}}{\sum_{h'=1}^H \widehat{\alpha}_{m,h'}^{\text{MAP}}} \frac{\widehat{\eta}_{l,h}^{\text{MAP}}}{\sum_{l'=1}^L \widehat{\eta}_{l',h}^{\text{MAP}}} \right). \end{aligned} \quad (16.88)$$

Let us first consider the case when $\alpha < 1$. In this case, F diverges to $F \rightarrow -\infty$ with fixed N , when $\widehat{\alpha}_{m,h} = O(1)$ for any (h, m) and $\widehat{\alpha}_{m,h'} \rightarrow +0$ for

all other $h' \neq h$. Therefore, the solution is sparse (so sparse that the estimator is useless). Similarly, assume that $\eta < 1$. Then F diverges to $F \rightarrow -\infty$ with fixed N , when $\widehat{\eta}_{l,h} = O(1)$ for any (l, h) and $\widehat{\eta}_{l',h} \rightarrow +0$ for all other $l' \neq l$. Therefore, the solution is useless.

When $\alpha \geq 1$ and $\eta \geq 1$, the solution satisfies the following stationary condition:

$$\widehat{\alpha}_{m,h}^{\text{MAP}} = \alpha - 1 + \sum_{n=1}^{N^{(m)}} \widehat{z}_h^{\text{MAP}(n,m)}, \quad \widehat{\eta}_{l,h}^{\text{MAP}} = \eta - 1 + \sum_{m=1}^M \sum_{n=1}^{N^{(m)}} w_l^{(n,m)} \widehat{z}_h^{\text{MAP}(n,m)}, \quad (16.89)$$

$$\widehat{z}_h^{\text{MAP}(n,m)} = \frac{\widehat{\alpha}_{m,h}^{\text{MAP}} \prod_{l=1}^L (\widehat{\eta}_{l,h}^{\text{MAP}})^{w_l^{(n,m)}}}{\sum_{h'=1}^H (\widehat{\alpha}_{m,h'}^{\text{MAP}} \prod_{l=1}^L (\widehat{\eta}_{l,h'}^{\text{MAP}})^{w_l^{(n,m)}})}. \quad (16.90)$$

In the same way as for VB, PB-A, and PB-B learning, we can obtain the following lemma:

Lemma 16.22 *Let $\widehat{\mathbf{B}}^{\text{MAP}} \widehat{\boldsymbol{\Theta}}^{\text{MAP}\top} = \langle \mathbf{B} \boldsymbol{\Theta}^\top \rangle_{r^{\text{MAP}}(\boldsymbol{\theta}, \mathbf{B})}$. Then Q^{MAP} is minimized when $\widehat{J} = O_p(N^{-1})$, and it holds that*

$$Q^{\text{MAP}} = NS_N(\mathcal{D}) + O_p(\widehat{J}N + LM).$$

R^{MAP} is written as follows:

$$R^{\text{MAP}} = \left\{ MH(\alpha - 1) + \widehat{H}L(\eta - 1) - \sum_{h=1}^{\widehat{H}} (\widehat{M}^{(h)}(\alpha - 1) + \widehat{L}^{(h)}(\eta - 1)) \right\} \log N + (H - \widehat{H})L(\eta - 1) \log L + O_p(H(M + L)). \quad (16.91)$$

Taking the different asymptotic limits, we obtain the following theorem:

Theorem 16.23 *When $\alpha < 1$, each row vector of $\widehat{\boldsymbol{\Theta}}^{\text{MAP}}$ has only one nonzero entry, and the MAP solution is sparse. When $\eta < 1$, each column vector of $\widehat{\mathbf{B}}^{\text{MAP}}$ has only one nonzero entry. Assume in the following that $\alpha, \eta \geq 1$. In the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$, it holds that $\widehat{J} = O_p(1/N)$ and*

$$\widetilde{F}^{\text{MAP}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{MAP}} \log N + O_p(1),$$

where

$$\lambda'_{\text{LDA}}^{\text{MAP}} = MH(\alpha - 1) + \widehat{H}L(\eta - 1) - \sum_{h=1}^{\widehat{H}} (\widehat{M}^{(h)}(\alpha - 1) + \widehat{L}^{(h)}(\eta - 1)).$$

In the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$, it holds that $\widehat{J} = o_p(\log N)$, and

$$\widetilde{F}^{\text{MAP}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{MAP}} \log N + o_p(N \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{MAP}} = MH(\alpha - 1) - \sum_{h=1}^{\widehat{H}} \widehat{M}^{(h)}(\alpha - 1).$$

In the limit when $N, L \rightarrow \infty$ with $\frac{L}{N}, M \sim O(1)$, it holds that $\widehat{J} = o_p(\log N)$, and

$$\widetilde{F}^{\text{MAP}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{MAP}} \log N + o_p(N \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{MAP}} = HL(\eta - 1).$$

In the limit when $N, L, M \rightarrow \infty$ with $\frac{L}{N}, \frac{M}{N} \sim O(1)$, it holds that $\widehat{J} = o_p(N \log N)$, and

$$\widetilde{F}^{\text{MAP}}(\mathcal{D}) = \lambda'_{\text{LDA}}^{\text{MAP}} \log N + o_p(N^2 \log N),$$

where

$$\lambda'_{\text{LDA}}^{\text{MAP}} = H(M(\alpha - 1) + L(\eta - 1)).$$

Theorem 16.23 states that the MAP solution is sparse when $\alpha < 1$. However, it provides no information on the sparsity of the MAP solution for $\eta < 1$. In the following, we investigate the sparsity of the solution for $\alpha, \eta \geq 1$.

In the Limit When $N \rightarrow \infty$ with $L, M \sim O(1)$

The coefficient of the leading term of the free energy is

$$\lambda'_{\text{LDA}}^{\text{MAP}} = MH(\alpha - 1) + \sum_{h=1}^{\widehat{H}} (L(\eta - 1) - \widehat{M}^{(h)}(\alpha - 1) - \widehat{L}^{(h)}(\eta - 1)).$$

The solution is sparse if $\lambda'_{\text{LDA}}^{\text{MAP}}$ is increasing with respect to \widehat{H} , and dense if it is decreasing. We focus on the case where $\alpha, \eta \geq 1$. Eqs. (16.57) through (16.59) imply the following:

The solution is sparse if

$$L(\eta - 1) - \max_h \left(M^{*(h)} (\alpha - 1) + L^{*(h)} (\eta - 1) \right) > 0, \quad (16.92)$$

and dense if

$$L(\eta - 1) - \max_h \left(M^{*(h)} (\alpha - 1) + L^{*(h)} (\eta - 1) \right) < 0. \quad (16.93)$$

Therefore, the solution is sparse if

$$\begin{aligned} L(\eta - 1) - \max_h M^{*(h)} (\alpha - 1) - \max_h L^{*(h)} (\eta - 1) &> 0, \text{ or equivalently,} \\ \alpha &< 1 + \frac{(L - \max_h L^{*(h)}) (\eta - 1)}{\max_h M^{*(h)}}, \end{aligned}$$

and dense if

$$\begin{aligned} L(\eta - 1) - \min_h M^{*(h)} (\alpha - 1) - \max_h L^{*(h)} (\eta - 1) &< 0, \text{ or equivalently,} \\ \alpha &> 1 + \frac{(L - \max_h L^{*(h)}) (\eta - 1)}{\min_h M^{*(h)}}. \end{aligned}$$

Thus, we can conclude that the solution is sparse if

$$\alpha < 1,$$

and dense if

$$\alpha > 1 + \frac{L(\eta - 1)}{\min_h M^{*(h)}}.$$

Summarizing the preceding, we have the following lemma:

Lemma 16.24 *Assume that $\eta \geq 1$. The solution is sparse if $\alpha < 1$, and dense if $\alpha > 1 + \frac{L(\eta - 1)}{\min_h M^{*(h)}}$.*

In the Limit When $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$

The coefficient of the leading term of the free energy is given by

$$\lambda_{\text{LDA}}^{\text{MAP}} = MH(\alpha - 1) - \sum_{h=1}^{\widehat{H}} \widehat{M}^{(h)}(\alpha - 1). \quad (16.94)$$

Although the predictive distribution does not necessarily converge to the true distribution, we can investigate the sparsity of the solution by considering the duplication rules (16.57) through (16.59) that keep $\widehat{\mathbf{B}}\boldsymbol{\theta}^\top$ unchanged. It is clear that Eq. (16.94) is decreasing with respect to \widehat{H} if $\alpha > 1$. Combing this result with Theorem 16.23, which states that the MAP solution is sparse if $\alpha < 1$, and Lemma 16.24, we obtain the following corollary:

Corollary 16.25 *Assume that $\eta \geq 1$. In the limit when $N \rightarrow \infty$ with $L, M \sim O(1)$, the MAP solution is sparse if $\alpha < 1$, and dense if $\alpha > 1 + \frac{L(\eta-1)}{\min_h M^{*(h)}}$. In the limit when $N, M \rightarrow \infty$ with $\frac{M}{N}, L \sim O(1)$, the MAP solution is sparse if $\alpha < 1$, and dense if $\alpha > 1$.*

Summary of Results

Summarizing Corollaries 16.11, 16.17, 16.21, and 16.25 completes the proof of Theorem 16.13. \square

17

Unified Theory for Latent Variable Models

In this chapter, we present a formula for evaluating an asymptotic form of the VB free energy of a general class of latent variable models by relating it to the asymptotic theory of Bayesian learning (Watanabe, 2012). This formula is applicable to all latent variable models discussed in Chapters 15 and 16.¹ It also explains relationships between these asymptotic analyses of VB free energy and several previous works where the asymptotic Bayes free energy has been analyzed for specific latent variable models. We apply this formula to Gaussian mixture models (GMMs) as an example and demonstrate another proof of the upper-bound of the VB free energy given in Section 15.2. Furthermore, this analysis also provides a quantity that is related to the generalization performance of VB learning. Analysis of generalization performance of VB learning has been conducted only for limited cases, as discussed in Chapter 14. We show inequalities that relate the VB free energy to the generalization errors of an approximate predictive distribution (Watanabe, 2012).

17.1 Local Latent Variable Model

Consider the joint model

$$p(\mathbf{x}, \mathbf{z}|\mathbf{w}) \tag{17.1}$$

on the observed variable \mathbf{x} and the local latent variable \mathbf{z} with the parameter \mathbf{w} . The marginal distribution of the observed variable is²

¹ The reduced rank regression (RRR) model discussed in Chapter 14 is not included in this class of latent variable models.

² The model is denoted as if the local latent variable is discrete, it can also be continuous. In this case, the sum \sum_z is replaced by the integral $\int dz$. The probabilistic principal component analysis is an example with a continuous local latent variable.

$$p(\mathbf{x}|\mathbf{w}) = \sum_z p(\mathbf{x}, z|\mathbf{w}). \quad (17.2)$$

For the complete data set $\{\mathcal{D}, \mathcal{H}\} = \{(\mathbf{x}^{(n)}, \mathbf{z}^{(1)}), \dots, (\mathbf{x}^{(N)}, \mathbf{z}^{(N)})\}$, we assume the i.i.d. model

$$p(\mathcal{D}, \mathcal{H}|\mathbf{w}) = \prod_{n=1}^N p(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}|\mathbf{w}),$$

which implies

$$\begin{aligned} p(\mathcal{D}|\mathbf{w}) &= \prod_{n=1}^N p(\mathbf{x}^{(n)}|\mathbf{w}), \\ p(\mathcal{H}|\mathcal{D}, \mathbf{w}) &= \prod_{n=1}^N p(\mathbf{z}^{(n)}|\mathbf{x}^{(n)}, \mathbf{w}). \end{aligned}$$

We assume that

$$p(\mathbf{x}|\mathbf{w}^*) = \sum_z p(\mathbf{x}, z|\mathbf{w}^*)$$

with the parameter \mathbf{w}^* is the underlying distribution generating data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. Because of the nonidentifiability of the latent variable model, the set of true parameters,

$$\mathcal{W}^* \equiv \left\{ \tilde{\mathbf{w}}^*; \sum_z p(\mathbf{x}, z|\tilde{\mathbf{w}}^*) = p(\mathbf{x}|\mathbf{w}^*) \right\}, \quad (17.3)$$

is not generally a point but can be a union of several manifolds with singularities as demonstrated in Section 13.5.

In the analysis in this chapter, we define and analyze quantities related to generalization performance of a *joint* model, where the local latent variables are treated as observed variables. Although we do not consider the case where the local latent variables are observed, those quantities are useful for relating generalization properties of VB learning to those of Bayesian learning, with which we establish a unified theory connecting VB learning and Bayesian learning of latent variable models.

Thus, consider for a moment the Bayesian learning of the joint model (17.1), where the complete data set $\{\mathcal{D}, \mathcal{H}\}$ is observed. For the prior distribution $p(\mathbf{w})$, the posterior distribution is given by

$$p(\mathbf{w}|\mathcal{D}, \mathcal{H}) = \frac{p(\mathcal{D}, \mathcal{H}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}, \mathcal{H})}. \quad (17.4)$$

The Bayes free energy of the joint model is defined by

$$F_{\text{Joint}}^{\text{Bayes}}(\mathcal{D}, \mathcal{H}) = -\log p(\mathcal{D}, \mathcal{H}) = -\log \int p(\mathcal{D}, \mathcal{H}|\mathbf{w})p(\mathbf{w})d\mathbf{w}.$$

If $\tilde{\mathbf{w}}^* \in \mathcal{W}^*$ is the true parameter, i.e., the complete data set $\{\mathcal{D}, \mathcal{H}\}$ is generated from $q(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)$ i.i.d., the relative Bayes free energy is defined by

$$\tilde{F}_{\text{Joint}}^{\text{Bayes}}(\mathcal{D}, \mathcal{H}) = F_{\text{Joint}}^{\text{Bayes}}(\mathcal{D}, \mathcal{H}) - S_N(\mathcal{D}, \mathcal{H}), \quad (17.5)$$

where

$$S_N(\mathcal{D}, \mathcal{H}) = -\frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}|\tilde{\mathbf{w}}^*)$$

is the empirical joint entropy. Then the average relative Bayes free energy is defined by

$$\overline{F}_{\text{Joint}}^{\text{Bayes}}(N) = \langle \tilde{F}_{\text{Joint}}^{\text{Bayes}}(\mathcal{D}, \mathcal{H}) \rangle_{p(\mathcal{D}, \mathcal{H}|\tilde{\mathbf{w}}^*)},$$

and the average Bayes generalization error of the predictive distribution for the joint model is defined by

$$\overline{\text{GE}}_{\text{Joint}}^{\text{Bayes}}(N) = \langle \text{KL}(p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)||p(\mathbf{x}, \mathbf{z}|\mathcal{D}, \mathcal{H})) \rangle_{p(\mathcal{D}, \mathcal{H}|\tilde{\mathbf{w}}^*)},$$

where

$$p(\mathbf{x}, \mathbf{z}|\mathcal{D}, \mathcal{H}) = \int p(\mathbf{x}, \mathbf{z}|\mathbf{w})p(\mathbf{w}|\mathcal{D}, \mathcal{H})d\mathbf{w}.$$

These two quantities are related to each other as Eq. (13.24):

$$\overline{\text{GE}}_{\text{Joint}}^{\text{Bayes}}(N) = \overline{F}_{\text{Joint}}^{\text{Bayes}}(N+1) - \overline{F}_{\text{Joint}}^{\text{Bayes}}(N). \quad (17.6)$$

Furthermore, the average relative Bayes free energy for the joint model can be approximated as (see Eq. (13.118))

$$\overline{F}_{\text{Joint}}^{\text{Bayes}}(N) \approx -\log \int \exp(-N\overline{E}(\mathbf{w})) \cdot p(\mathbf{w})d\mathbf{w}, \quad (17.7)$$

where

$$\overline{E}(\mathbf{w}) = \text{KL}(p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)||p(\mathbf{x}, \mathbf{z}|\mathbf{w})) = \left\langle \log \frac{p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)}{p(\mathbf{x}, \mathbf{z}|\mathbf{w})} \right\rangle_{p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)}. \quad (17.8)$$

Since the log-sum inequality yields that³

$$\sum_z p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*) \log \frac{p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)}{p(\mathbf{x}, \mathbf{z}|\mathbf{w})} \geq p(\mathbf{x}|\mathbf{w}^*) \log \frac{p(\mathbf{x}|\mathbf{w}^*)}{p(\mathbf{x}|\mathbf{w})},$$

we have

$$\bar{E}(\mathbf{w}) \geq E(\mathbf{w}), \quad (17.10)$$

where

$$E(\mathbf{w}) = \text{KL}(p(\mathbf{x}|\mathbf{w}^*)||p(\mathbf{x}|\mathbf{w})) = \left\langle \log \frac{p(\mathbf{x}|\mathbf{w}^*)}{p(\mathbf{x}|\mathbf{w})} \right\rangle_{p(\mathbf{x}|\mathbf{w}^*)}.$$

Hence, it follows from Eq. (13.118) that

$$\begin{aligned} \bar{F}^{\text{Bayes}}(N) &= \left\langle \bar{F}^{\text{Bayes}}(\mathcal{D}) \right\rangle_{q(\mathcal{D})} \\ &\approx -\log \int \exp(-NE(\mathbf{w})) \cdot p(\mathbf{w}) d\mathbf{w} \\ &\leq -\log \int \exp(-N\bar{E}(\mathbf{w})) \cdot p(\mathbf{w}) d\mathbf{w} \approx \bar{F}_{\text{Joint}}^{\text{Bayes}}(N), \end{aligned} \quad (17.11)$$

where

$$\bar{F}^{\text{Bayes}}(\mathcal{D}) = F^{\text{Bayes}}(\mathcal{D}) - NS_N(\mathcal{D})$$

is the relative Bayes free energy defined by the Bayes free energy of the original marginal model,

$$F^{\text{Bayes}}(\mathcal{D}) = -\log p(\mathcal{D}) = -\log \int p(\mathcal{D}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$

and its empirical entropy,

$$S_N(\mathcal{D}) = -\frac{1}{N} \log p(\mathcal{D}|\mathbf{w}^*) = -\frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}^{(n)}|\mathbf{w}^*), \quad (17.12)$$

as in Section 13.3.2.

The asymptotic theory of Bayesian learning (Theorem 13.13) shows that an asymptotic form of $\bar{F}_{\text{Joint}}^{\text{Bayes}}(N)$ is given by

$$\bar{F}_{\text{Joint}}^{\text{Bayes}}(N) = \lambda_{\text{Joint}}^{\text{Bayes}} \log N - (m_{\text{Joint}}^{\text{Bayes}} - 1) \log \log N + O(1), \quad (17.13)$$

³ The log-sum inequality is the following inequality satisfied for nonnegative reals $a_i \geq 0$ and $b_i \geq 0$:

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i \right) \log \frac{(\sum_i a_i)}{(\sum_i b_i)}. \quad (17.9)$$

This can be proved by subtracting the right-hand side from the left-hand side and applying the nonnegativity of the KL divergence.

where $-\lambda'_{\text{Joint}}^{\text{Bayes}}$ and $m'_{\text{Joint}}^{\text{Bayes}}$ are respectively the largest pole and its order of the zeta function defined for a complex number z by

$$\zeta_{\bar{E}}(z) = \int \bar{E}(\mathbf{w})^z p(\mathbf{w}) d\mathbf{w}. \quad (17.14)$$

This means that the asymptotic behavior of the free energy is characterized by $\bar{E}(\mathbf{w})$, while that of the Bayes free energy F^{Bayes} is characterized by $E(\mathbf{w}) = \text{KL}(p(\mathbf{x}|\mathbf{w}^*)||p(\mathbf{x}|\mathbf{w}))$ and the zeta function $\zeta_E(z)$ in Eq. (13.122) as Theorem 13.13. The two functions, E and \bar{E} , are related by the log-sum inequality (17.10).

Then Corollary 13.14 implies the following asymptotic expansion of the average generalization error:

$$\overline{\text{GE}}_{\text{Joint}}^{\text{Bayes}}(N) = \frac{\lambda'_{\text{Joint}}^{\text{Bayes}}}{N} - \frac{m'_{\text{Joint}}^{\text{Bayes}} - 1}{N \log N} + o\left(\frac{1}{N \log N}\right). \quad (17.15)$$

With the preceding quantities, we first provide a general upper-bound for the VB free energy (Section 17.2), and then show inequalities relating the VB free energy to the generalization errors of an approximate predictive distribution for the joint model (Section 17.4).

17.2 Asymptotic Upper-Bound for VB Free Energy

Given the training data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$, consider VB learning for the latent variable model (17.2) with the prior distribution $p(\mathbf{w})$. Under the constraint,

$$r(\mathbf{w}, \mathcal{H}) = r_w(\mathbf{w})r_{\mathcal{H}}(\mathcal{H}),$$

the VB free energy is defined by

$$F^{\text{VB}}(\mathcal{D}) = \min_{r_w(\mathbf{w}), r_{\mathcal{H}}(\mathcal{H})} F(r),$$

where

$$F(r) = \left\langle \log \frac{r_w(\mathbf{w})r_{\mathcal{H}}(\mathcal{H})}{p(\mathcal{D}, \mathcal{H}|\mathbf{w})p(\mathbf{w})} \right\rangle_{r_w(\mathbf{w})r_{\mathcal{H}}(\mathcal{H})} \quad (17.16)$$

$$= F^{\text{Bayes}}(\mathcal{D}) + \text{KL}(r_w(\mathbf{w})r_{\mathcal{H}}(\mathcal{H})||p(\mathbf{w}, \mathcal{H}|\mathcal{D})). \quad (17.17)$$

The stationary condition of the free energy yields

$$r_w(\mathbf{w}) = \frac{1}{C_w} p(\mathbf{w}) \exp \langle \log p(\mathcal{D}, \mathcal{H}|\mathbf{w}) \rangle_{r_{\mathcal{H}}(\mathcal{H})}, \quad (17.18)$$

$$r_{\mathcal{H}}(\mathcal{H}) = \frac{1}{C_{\mathcal{H}}} \exp \langle \log p(\mathcal{D}, \mathcal{H}|\mathbf{w}) \rangle_{r_w(\mathbf{w})}. \quad (17.19)$$

Let us define the relative VB free energy

$$\bar{F}^{\text{VB}}(\mathcal{D}) = F^{\text{VB}}(\mathcal{D}) - NS_N(\mathcal{D})$$

by the VB free energy and the empirical entropy (17.12). For arbitrary $\tilde{\mathbf{w}}^* \in \mathcal{W}^*$, substituting Eq. (17.18) into Eq. (17.16), we have

$$\begin{aligned} \bar{F}^{\text{VB}}(\mathcal{D}) &= \min_{r_{\mathcal{H}}(\mathcal{H})} \left[-\log \int p(\mathbf{w}) \exp \left\langle \log \frac{p(\mathcal{D}, \mathcal{H}|\mathbf{w})}{r_{\mathcal{H}}(\mathcal{H})} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})} d\mathbf{w} \right] + \log p(\mathcal{D}|\mathbf{w}^*) \\ &\quad (17.20) \end{aligned}$$

$$\begin{aligned} &\leq -\log \int \exp \left\{ \sum_{\mathcal{H}} p(\mathcal{H}|\mathcal{D}, \tilde{\mathbf{w}}^*) \log \frac{p(\mathcal{D}, \mathcal{H}|\mathbf{w})}{p(\mathcal{D}, \mathcal{H}|\tilde{\mathbf{w}}^*)} \right\} p(\mathbf{w}) d\mathbf{w} \quad (17.21) \\ &\equiv \bar{F}^{\text{VB}*}(\mathcal{D}). \end{aligned}$$

Here, we have substituted $r_{\mathcal{H}}(\mathcal{H}) \leftarrow p(\mathcal{H}|\mathcal{D}, \tilde{\mathbf{w}}^*) = \frac{p(\mathcal{D}, \mathcal{H}|\tilde{\mathbf{w}}^*)}{\sum_{\mathcal{H}} p(\mathcal{D}, \mathcal{H}|\tilde{\mathbf{w}}^*)}$ to obtain the upper-bound (17.21). The expression (17.20) of the free energy corresponds to viewing the VB learning as a local variational approximation (Section 5.3.3), where the variational parameter $\mathbf{h}(\xi)$ is the vector consisting of the elements $\log p(\mathcal{D}, \mathcal{H}, \xi)$ for all possible \mathcal{H} .⁴

By taking the expectation with respect to the distribution of training samples, we define the average relative VB free energy and its upper-bound as

$$\bar{F}^{\text{VB}}(N) = \left\langle \bar{F}^{\text{VB}}(\mathcal{D}) \right\rangle_{p(\mathcal{D}|\mathbf{w}^*)}, \quad (17.22)$$

$$\bar{F}^{\text{VB}*}(N) = \left\langle \bar{F}^{\text{VB}*}(\mathcal{D}) \right\rangle_{p(\mathcal{D}|\mathbf{w}^*)}. \quad (17.23)$$

From Eq. (17.7), we have

$$\bar{F}_{\text{Joint}}^{\text{Bayes}}(N) \approx -\log \int e^{-N\bar{E}(\mathbf{w})} p(\mathbf{w}) d\mathbf{w} \equiv \bar{F}_{\text{Joint}}^{\text{Bayes}}(N),$$

where $\bar{E}(\mathbf{w})$ is defined by Eq. (17.8). Then, the following theorem holds:

Theorem 17.1 *It holds that*

$$\bar{F}_{\text{Joint}}^{\text{Bayes}}(N) \leq \bar{F}^{\text{VB}}(N) \leq \bar{F}^{\text{VB}*}(N) \leq \bar{F}_{\text{Joint}}^{\text{Bayes}}(N). \quad (17.24)$$

Proof The left inequality follows from Eq. (17.17). Eq. (17.21) gives

$$\begin{aligned} \bar{F}^{\text{VB}}(N) &\leq \bar{F}^{\text{VB}*}(N) \\ &= \left\langle \bar{F}^{\text{VB}*}(\mathcal{D}) \right\rangle_{p(\mathcal{D}|\mathbf{w}^*)} \end{aligned}$$

⁴ The variational parameter $\mathbf{h}(\xi)$ has one-to-one correspondence with $p(\mathcal{H}|\mathcal{D}, \xi)$, and is substituted as $\mathbf{h}(\xi) \leftarrow \mathbf{h}(\tilde{\mathbf{w}}^*)$ in Eq. (17.21).

$$\begin{aligned}
&= - \left\langle \log \int \exp \left\{ \sum_{\mathcal{H}} p(\mathcal{H}|\mathcal{D}, \tilde{\mathbf{w}}^*) \log \frac{p(\mathcal{D}, \mathcal{H}|\mathbf{w})}{p(\mathcal{D}, \mathcal{H}|\tilde{\mathbf{w}}^*)} \right\} p(\mathbf{w}) d\mathbf{w} \right\rangle_{p(\mathcal{D}|\mathbf{w}^*)} \\
&\leq - \log \int \exp \left\{ \left\langle \sum_{\mathcal{H}} p(\mathcal{H}|\mathcal{D}, \tilde{\mathbf{w}}^*) \log \frac{p(\mathcal{D}, \mathcal{H}|\mathbf{w})}{p(\mathcal{D}, \mathcal{H}|\tilde{\mathbf{w}}^*)} \right\rangle_{p(\mathcal{D}|\mathbf{w}^*)} \right\} p(\mathbf{w}) d\mathbf{w} \\
&= - \log \int e^{-N\bar{E}(\mathbf{w})} p(\mathbf{w}) d\mathbf{w} = F_{\text{Joint}}^{\text{Bayes}}(N).
\end{aligned}$$

The first and second equalities are definitions of $\bar{F}^{\text{VB}*}(N)$ and $\bar{F}^{\text{VB}*}(\mathcal{D})$. We have applied Jensen's inequality to the convex function $\log \int \exp(\cdot) p(\mathbf{w}) d\mathbf{w}$ to obtain the last inequality. Finally, the last equality follows from the fact that $p(\mathcal{D}|\mathbf{w}^*)p(\mathcal{H}|\mathcal{D}, \tilde{\mathbf{w}}^*) = p(\mathcal{D}, \mathcal{H}|\tilde{\mathbf{w}}^*)$ and the i.i.d. assumption. \square

The following corollary is immediately obtained from Theorems 13.13 and 17.1:

Corollary 17.2 *Let $0 > -\lambda_1 > -\lambda_2 > \dots$ be the sequence of the poles of the zeta function (17.14) in the decreasing order, and m_1, m_2, \dots be the corresponding orders of the poles. Then the average relative VB free energy (17.22) can be asymptotically upper-bounded as*

$$\bar{F}^{\text{VB}}(N) \leq \lambda_1 \log N - (m_1 - 1) \log \log N + O(1). \quad (17.25)$$

It holds in Eqs. (17.13) and (17.15) that $\lambda_{\text{Joint}}^{\text{Bayes}} = \lambda_1$ and $m_{\text{Joint}}^{\text{Bayes}} = m_1$ for λ_1 and m_1 defined in Corollary 17.2. Note that $\bar{E}(\mathbf{w})$ depends on $\tilde{\mathbf{w}}^* \in \mathcal{W}^*$. For different $\tilde{\mathbf{w}}^*$, we have different values of λ_1 , which is determined by the minimum over different $\tilde{\mathbf{w}}^* \in \mathcal{W}^*$ in Eq. (17.25). Then m_1 is determined by the maximum of the order of the pole for the minimum λ_1 . Also note that unlike for Bayesian learning, even if the largest pole of the zeta function is obtained, Eq. (17.25) does not necessarily provide a lower-bound of the VB free energy.

If the joint model $p(\mathbf{x}, \mathbf{z}|\mathbf{w})$, the true distribution $p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)$, and the prior $p(\mathbf{w})$ satisfy the regularity conditions (Section 13.4.1), it holds that

$$2\lambda_{\text{Joint}}^{\text{Bayes}} = D,$$

where D is the number of parameters.

If the joint model $p(\mathbf{x}, \mathbf{z}|\mathbf{w})$ is identifiable, even though the true parameter is on the boundary of the parameter space or the prior does not satisfy $0 < p(\mathbf{w}) < \infty$, $\lambda_{\text{Joint}}^{\text{Bayes}}$ can be analyzed similarly to the case of regular models. The GMM with redundant components is an example of such a case, as will be detailed in the next section.

If the joint model $p(\mathbf{x}, \mathbf{z}|\mathbf{w})$ is unidentifiable, we need the algebraic geometrical technique to analyze $\lambda'_{\text{Joint}}^{\text{Bayes}}$ as discussed in Section 13.5.4. This technique is also applicable to identifiable cases as will be demonstrated in a part of the analysis of $\lambda'_{\text{Joint}}^{\text{Bayes}}$ for the GMM in the last part of the next section.

17.3 Example: Average VB Free Energy of Gaussian Mixture Model

In this section, we derive an asymptotic upper-bound of the VB free energy of GMMs. Although this upper-bound is immediately obtained from Theorem 15.5 in Section 15.2, it was derived by direct evaluation and minimization of the free energy with respect to the expected sufficient statistics. In this section, we present another derivation through Theorem 17.1 in order to illustrate how the general theory described in Section 17.2 is applied.

Let

$$g(\mathbf{x}|\boldsymbol{\mu}) = \text{Gauss}_M(\mathbf{x}; \boldsymbol{\mu}, \mathbf{I}_M)$$

be the M -dimensional uncorrelated Gaussian density and consider the GMM with K components,

$$p(\mathbf{x}|\mathbf{w}) = \sum_z p(\mathbf{x}, \mathbf{z}|\mathbf{w}) = \sum_{k=1}^K \alpha_k g(\mathbf{x}|\boldsymbol{\mu}_k),$$

where $\mathbf{x} \in \mathbb{R}^M$ and $\mathbf{w} = (\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)$ denote the parameter vector consisting the mixing weights and the mean vectors, respectively.

Assuming the same prior given by

$$p(\boldsymbol{\alpha}|\boldsymbol{\phi}) = \text{Dirichlet}_K(\boldsymbol{\alpha}; (\boldsymbol{\phi}, \dots, \boldsymbol{\phi})^\top), \quad (17.26)$$

$$p(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, \xi) = \text{Gauss}_M(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, (1/\xi)\mathbf{I}_M), \quad (17.27)$$

and the same true distribution

$$q(\mathbf{x}) = p(\mathbf{x}|\mathbf{w}^*) = \sum_{k=1}^{K_0} \alpha_k^* g(\mathbf{x}|\boldsymbol{\mu}_k^*), \quad (17.28)$$

as in Sections 4.1.1 and 15.2, we immediately obtain from the upper-bound in Eq. (15.36) of Theorem 15.5 that

$$\bar{F}^{\text{VB}}(N) \leq \bar{\lambda}_{\text{MM}}^{\text{VB}} \log N + O(1), \quad (17.29)$$

where

$$\lambda'_{\text{MM}}^{\text{VB}} = \begin{cases} (K - K_0)\phi + \frac{MK_0 + K_0 - 1}{2} & (\phi < \frac{M+1}{2}), \\ \frac{MK + K - 1}{2} & (\phi \geq \frac{M+1}{2}). \end{cases}$$

In this section, we derive this upper-bound by using Theorem 17.1, which provides an alternative proof to the one presented in Section 15.2. Similar techniques were used for analyzing the Bayes free energy (13.19) in the asymptotic limit (Yamazaki and Watanabe, 2003a,b, 2005). Here, we evaluate the VB free energy and present the details of the proof for the specific choice of the prior distribution.

First, in order to define $p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)$ for \mathbf{z} with K elements, we extend and redefine the true parameter \mathbf{w}^* denoting it as $\tilde{\mathbf{w}}^* = (\tilde{\alpha}^*, [\tilde{\mu}_k^*]_{k=1}^K)$. Suppose that the true distribution with parameter $\tilde{\mathbf{w}}^*$ has \tilde{K} nonzero mixing weights. For example, we can assume that

$$\begin{aligned} \tilde{\alpha}_k^* &= \begin{cases} \alpha_k^* & (1 \leq k \leq K_0 - 1), \\ \alpha_{K_0}^*/(\tilde{K} - K_0 + 1) & (K_0 \leq k \leq \tilde{K}), \\ 0 & (\tilde{K} + 1 \leq k \leq K), \end{cases} \\ \tilde{\mu}_k^* &= \begin{cases} \mu_k^* & (1 \leq k \leq K_0), \\ \mu_{K_0}^* & (K_0 + 1 \leq k \leq K). \end{cases} \end{aligned}$$

Note that the marginal distribution of $p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)$ is reduced to Eq. (17.28). Then we have

$$\begin{aligned} \bar{E}(\mathbf{w}) &= \int \sum_z p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*) \log \frac{p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)}{p(\mathbf{x}, \mathbf{z}|\mathbf{w})} d\mathbf{x} \\ &= \int \sum_{k=1}^K \tilde{\alpha}_k^* g(\mathbf{x}|\tilde{\mu}_k^*) \log \frac{\tilde{\alpha}_k^* g(\mathbf{x}|\tilde{\mu}_k^*)}{\alpha_k g(\mathbf{x}|\mu_k)} d\mathbf{x} \\ &= \sum_{k=1}^K \tilde{\alpha}_k^* \left\{ \log \frac{\tilde{\alpha}_k^*}{\alpha_k} + \int g(\mathbf{x}|\tilde{\mu}_k^*) \log \frac{g(\mathbf{x}|\tilde{\mu}_k^*)}{g(\mathbf{x}|\mu_k)} d\mathbf{x} \right\} \\ &= \sum_{k=1}^{\tilde{K}} \tilde{\alpha}_k^* \left\{ \log \frac{\tilde{\alpha}_k^*}{\alpha_k} + \frac{\|\mu_k - \tilde{\mu}_k^*\|^2}{2} \right\}. \end{aligned}$$

Second, we divide the parameter \mathbf{w} into three parts,

$$\mathbf{w}_1 = (\alpha_2, \alpha_3, \dots, \alpha_{\tilde{K}}), \quad (17.30)$$

$$\mathbf{w}_2 = (\alpha_{\tilde{K}+1}, \dots, \alpha_K), \quad (17.31)$$

$$\mathbf{w}_3 = (\mu_1, \mu_2, \dots, \mu_{\tilde{K}}), \quad (17.32)$$

and define

$$\mathcal{W}_1 = \{\mathbf{w}_1; |\alpha_k - \tilde{\alpha}_k^*| \leq \epsilon, 2 \leq k \leq \hat{K}\},$$

$$\mathcal{W}_2 = \{\mathbf{w}_2; |\alpha_k| \leq \epsilon, \hat{K} \leq k \leq K\},$$

$$\mathcal{W}_3 = \{\mathbf{w}_3; \|\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k^*\| \leq \epsilon, 1 \leq k \leq \hat{K}\},$$

for a sufficiently small constant ϵ . For an arbitrary parameter $\mathbf{w} \in \mathcal{W}_1 \times \mathcal{W}_2 \times \mathcal{W}_3 \equiv \mathcal{W}(\epsilon)$, we can decompose $\bar{E}(\mathbf{w})$ as

$$\bar{E}(\mathbf{w}) = \bar{E}_1(\mathbf{w}_1) + \bar{E}_2(\mathbf{w}_2) + \bar{E}_3(\mathbf{w}_3), \quad (17.33)$$

where

$$\begin{aligned} \bar{E}_1(\mathbf{w}_1) &= \sum_{k=2}^{\hat{K}} \tilde{\alpha}_k^* \log \frac{\tilde{\alpha}_k^*}{\alpha_k} + \left(1 - \sum_{k=2}^{\hat{K}} \tilde{\alpha}_k^*\right) \log \frac{1 - \sum_{k=2}^{\hat{K}} \tilde{\alpha}_k^*}{1 - \sum_{k=2}^{\hat{K}} \alpha_k}, \\ \bar{E}_2(\mathbf{w}_2) &= \frac{1}{1-c} \frac{1 - \sum_{k=2}^{K_0} \tilde{\alpha}_k^*}{1 - \sum_{k=2}^{\hat{K}} \alpha_k} \sum_{k=\hat{K}+1}^K \alpha_k, \\ \bar{E}_3(\mathbf{w}_3) &= \sum_{k=1}^{\hat{K}} \frac{\tilde{\alpha}_k^*}{2} \|\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k^*\|^2. \end{aligned} \quad (17.34)$$

Here we have used the mean value theorem $-\log(1-t) = \frac{1}{1-c}t$ for some c , $0 \leq c \leq t$ with $t = \frac{\sum_{k=\hat{K}+1}^K \alpha_k}{1 - \sum_{k=2}^{\hat{K}} \alpha_k}$. Furthermore, for $\mathbf{w} \in \mathcal{W}(\epsilon)$, there exist positive constants C_1, C_2, C_3 , and C_4 such that

$$C_1 \sum_{k=2}^{\hat{K}} (\alpha_k - \tilde{\alpha}_k^*)^2 \leq \bar{E}_1(\mathbf{w}_1) \leq C_2 \sum_{k=2}^{\hat{K}} (\alpha_k - \tilde{\alpha}_k^*)^2, \quad (17.35)$$

$$C_3 \sum_{k=\hat{K}+1}^K \alpha_k \leq \bar{E}_2(\mathbf{w}_2) \leq C_4 \sum_{k=\hat{K}+1}^K \alpha_k. \quad (17.36)$$

Third, we evaluate the partial free energies defined for $i = 1, 2, 3$ by

$$F_i = -\log \int_{\mathcal{W}_i} \exp(-N\bar{E}_i(\mathbf{w}_i)) p(\mathbf{w}_i) d\mathbf{w}_i, \quad (17.37)$$

where $p(\mathbf{w}_i)$ is the product of factors of the prior in Eqs. (17.26) and (17.27), which involve \mathbf{w}_i defined in Eqs. (17.30) through (17.32).

It follows from Eqs. (17.24), (17.33), and (17.37) that

$$\bar{F}^{\text{VB}}(N) \leq F_1 + F_2 + F_3 + O(1). \quad (17.38)$$

From Eqs. (17.35) and (17.34), as for F_1 and F_3 , the Gaussian integration yields

$$F_1 = \frac{\widehat{K} - 1}{2} \log N + O(1), \quad (17.39)$$

$$F_3 = \frac{M\widehat{K}}{2} \log N + O(1). \quad (17.40)$$

Since

$$N^\phi \int_0^\epsilon e^{-n\alpha_k} \alpha_k^{\phi-1} d\alpha_k \rightarrow \Gamma(\phi) \quad (N \rightarrow \infty),$$

for $k = \widehat{K} + 1, \dots, K$, it follows from Eq. (17.36) that

$$F_2 = (K - \widehat{K})\phi \log N + O(1). \quad (17.41)$$

Finally, combining Eqs. (17.38) through (17.41), we obtain

$$\bar{F}^{\text{VB}}(N) \leq \left\{ (K - \widehat{K})\phi + \frac{M\widehat{K} + \widehat{K} - 1}{2} \right\} \log N + O(1).$$

Minimizing the right-hand side of the preceding expression over \widehat{K} ($K_0 \leq \widehat{K} \leq K$) leads to the upper-bound in Eq. (17.29).

Alternatively, the preceding evaluations of all the partial free energies, F_1 , F_2 , and F_3 , are obtained by using the algebraic geometrical method based on Corollary 17.2. For example, as for F_2 , the zeta function

$$\zeta_{\bar{E}_2}(z) = \int \bar{E}_2(\mathbf{w}_2)^z p(\mathbf{w}_2) d\mathbf{w}_2$$

has a pole $z = -(K - \widehat{K})\phi$. This can be observed by the change of variables, the so-called blow-up,

$$\begin{aligned} \alpha_k &= \alpha'_k \alpha'_K \quad (k = \widehat{K} + 1, \dots, K - 1), \\ \alpha_K &= \alpha'_K, \end{aligned}$$

which yields that $\zeta_{\bar{E}_2}$ has a term

$$\int \alpha'_K^z \alpha'_K^{(K-\widehat{K})\phi-1} \tilde{\zeta}_{\bar{E}_2}(\tilde{\mathbf{w}}'_2) d\alpha'_K = \frac{\tilde{\zeta}_{\bar{E}_2}(\tilde{\mathbf{w}}'_2)}{z + (K - \widehat{K})\phi},$$

where $\tilde{\zeta}_{\bar{E}_2}(\tilde{\mathbf{w}}'_2)$ is a function proportional to

$$\int \left(\sum_{k=\widehat{K}+1}^{K-1} \alpha'_k + 1 \right)^z \prod_{k=\widehat{K}+1}^{K-1} \alpha'_k^{\phi-1} \prod_{k=\widehat{K}+1}^{K-1} d\alpha'_k.$$

Hence, we can see that $\zeta_{\bar{E}_2}$ has a pole at $z = -(K - \widehat{K})\phi$.

17.4 Free Energy and Generalization Error

In this section, we relate the VB free energy to the generalization performance of VB learning. We denote a training data set by $\mathcal{D}^N = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\}$ with the number N of training samples as a superscript in this section.

Let $p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*)$ be the true distribution of the observed variable \mathbf{x} and the latent variable \mathbf{z} , which has the marginal distribution $p(\mathbf{x}|\mathbf{w}^*)$. We define the generalization error of the predictive distribution for the *joint distribution*,

$$p^{\text{VB}*}(\mathbf{x}, \mathbf{z}|\mathcal{D}^N) = \langle p(\mathbf{x}, \mathbf{z}|\mathbf{w}) \rangle_{r^*(\mathbf{w}; \tilde{\mathbf{w}}^*)} = \int p(\mathbf{x}, \mathbf{z}|\mathbf{w}) r^*(\mathbf{w}; \tilde{\mathbf{w}}^*) d\mathbf{w}, \quad (17.42)$$

by the Bayes generalization error (13.133)

$$\text{GE}_{\text{Joint}}^{\text{VB}*}(\mathcal{D}^N) = \text{KL}(p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*) \| p^{\text{VB}*}(\mathbf{x}, \mathbf{z}|\mathcal{D}^N)), \quad (17.43)$$

and the Gibbs generalization error (13.136) by

$$\text{GGE}_{\text{Joint}}^{\text{VB}*}(\mathcal{D}^N) = \langle \text{KL}(p(\mathbf{x}, \mathbf{z}|\tilde{\mathbf{w}}^*) \| p(\mathbf{x}, \mathbf{z}|\mathbf{w})) \rangle_{r^*(\mathbf{w}; \tilde{\mathbf{w}}^*)}, \quad (17.44)$$

where

$$r^*(\mathbf{w}; \tilde{\mathbf{w}}^*) \propto p(\mathbf{w}) \prod_{n=1}^N \exp \left(- \sum_z p(z|\mathbf{x}^{(n)}, \tilde{\mathbf{w}}^*) \log \frac{p(\mathbf{x}^{(n)}, z|\tilde{\mathbf{w}}^*)}{p(\mathbf{x}^{(n)}, z|\mathbf{w})} \right)$$

is the approximate posterior distribution (17.18) with $p(\mathcal{H}|\mathcal{D}^N, \tilde{\mathbf{w}}^*)$ substituted for $r_{\mathcal{H}}(\mathcal{H})$. We denote their means by

$$\begin{aligned} \overline{\text{GE}}_{\text{Joint}}^{\text{VB}*}(N) &= \langle \text{GE}_{\text{Joint}}^{\text{VB}*}(\mathcal{D}^N) \rangle_{p(\mathcal{D}^N|\mathbf{w}^*)}, \\ \overline{\text{GGE}}_{\text{Joint}}^{\text{VB}*}(N) &= \langle \text{GGE}_{\text{Joint}}^{\text{VB}*}(\mathcal{D}^N) \rangle_{p(\mathcal{D}^N|\mathbf{w}^*)}. \end{aligned}$$

Then the following theorem holds:

Theorem 17.3 *It holds that*

$$\overline{\text{GE}}_{\text{Joint}}^{\text{VB}*}(N) \leq \overline{F}^{\text{VB}*}(N+1) - \overline{F}^{\text{VB}*}(N) \leq \overline{\text{GGE}}_{\text{Joint}}^{\text{VB}*}(N), \quad (17.45)$$

where $\overline{F}^{\text{VB}*}(N)$ is the upper-bound (17.23) of the average relative VB free energy.

Proof We have

$$\begin{aligned} &\widetilde{F}^{\text{VB}*}(\mathcal{D}^{N+1}) - \widetilde{F}^{\text{VB}*}(\mathcal{D}^N) \\ &= -\log \frac{\int \prod_{n=1}^{N+1} \exp \left(\sum_z p(z|\mathbf{x}^{(n)}, \tilde{\mathbf{w}}^*) \log \frac{p(\mathbf{x}^{(n)}, z|\mathbf{w})}{p(\mathbf{x}^{(n)}, z|\tilde{\mathbf{w}}^*)} \right) p(\mathbf{w}) d\mathbf{w}}{\int \prod_{n=1}^N \exp \left(\sum_z p(z|\mathbf{x}^{(n)}, \tilde{\mathbf{w}}^*) \log \frac{p(\mathbf{x}^{(n)}, z|\mathbf{w})}{p(\mathbf{x}^{(n)}, z|\tilde{\mathbf{w}}^*)} \right) p(\mathbf{w}) d\mathbf{w}} \end{aligned}$$

$$\begin{aligned}
&= -\log \int \exp \left(\sum_z p(z|x^{(N+1)}, \tilde{\mathbf{w}}^*) \log \frac{p(\mathbf{x}^{(N+1)}, z|\mathbf{w})}{p(\mathbf{x}^{(N+1)}, z|\tilde{\mathbf{w}}^*)} \right) r^*(\mathbf{w}; \tilde{\mathbf{w}}^*) d\mathbf{w} \\
&= \sum_z p(z|x^{(N+1)}, \tilde{\mathbf{w}}^*) \log p(\mathbf{x}^{(N+1)}, z|\tilde{\mathbf{w}}^*) \\
&\quad - \log \int \exp \left(\langle \log p(\mathbf{x}^{(N+1)}, z|\mathbf{w}) \rangle_{p(z|x^{(N+1)}, \tilde{\mathbf{w}}^*)} \right) r^*(\mathbf{w}; \tilde{\mathbf{w}}^*) d\mathbf{w} \quad (17.46) \\
&\geq \sum_z p(z|x^{(N+1)}, \tilde{\mathbf{w}}^*) \log \frac{p(\mathbf{x}^{(N+1)}, z|\tilde{\mathbf{w}}^*)}{\langle p(\mathbf{x}^{(N+1)}, z|\mathbf{w}) \rangle_{r^*(\mathbf{w}; \tilde{\mathbf{w}}^*)}}.
\end{aligned}$$

In the last inequality, we have applied Jensen's inequality to the convex function $\log \int \exp(\cdot) p(\mathbf{w}) d\mathbf{w}$. Taking the expectation with respect to $\prod_{n=1}^{N+1} p(\mathbf{x}^{(n)}|\mathbf{w}^*)$ in both sides of the preceding inequality yields the left inequality in Eq. (17.45).

By applying Jensen's inequality for the exponential function in Eq. (17.46), and taking the expectation, we have the right inequality in Eq. (17.45). \square

The inequalities in Eq. (17.45) are analogous to Eq. (17.6). Let $\lambda'^{\text{VB}*}$ be the free energy coefficient of $\bar{F}^{\text{VB}*}(N)$, i.e.,

$$\bar{F}^{\text{VB}*}(N) = \lambda'^{\text{VB}*} \log N + o(\log N). \quad (17.47)$$

If its difference has the asymptotic form

$$\bar{F}^{\text{VB}*}(N+1) - \bar{F}^{\text{VB}*}(N) = \frac{\lambda'^{\text{VB}*}}{N} + o\left(\frac{1}{N}\right),$$

the left inequality in Eq. (17.45) suggests that

$$\overline{\text{GE}}_{\text{Joint}}^{\text{VB}*}(N) \leq \frac{\lambda'^{\text{VB}*}}{N} + o\left(\frac{1}{N}\right).$$

This means that the free energy coefficient $\lambda'^{\text{VB}*}$ of $\bar{F}^{\text{VB}*}(N)$ is directly related to the generalization error of VB learning measured by Eq. (17.43). Theorem 17.1 implies that the free energy coefficients satisfy $\lambda'^{\text{VB}*} \leq \lambda'^{\text{Bayes}}_{\text{Joint}}$, which in turn implies from Eq. (17.6) that

$$\overline{\text{GE}}_{\text{Joint}}^{\text{VB}*}(N) \leq \overline{\text{GE}}_{\text{Joint}}^{\text{Bayes}}(N) \quad (17.48)$$

holds asymptotically.

Let

$$\widehat{r}_{\mathbf{w}}(\mathbf{w}) = \operatorname{argmin}_{r_{\mathbf{w}}(\mathbf{w})} \min_{r_{\mathcal{H}}(\mathcal{H})} F(r)$$

be the optimal VB posterior of the parameter that minimizes the free energy (17.16). The average generalization errors of VB learning are naturally defined by

$$\overline{GE}_{Joint}^{VB}(N) = \left\langle KL(p(x, z|\tilde{w}^*) || p^{VB}(x, z|\mathcal{D}^N)) \right\rangle_{q(\mathcal{D}^N)}$$

for the joint predictive distribution $p^{VB}(x, z|\mathcal{D}^N) = \langle p(x, z|w) \rangle_{\widehat{r}_w(w)}$, and by

$$\overline{GE}^{VB}(N) = \left\langle KL(p(x|w^*) || p^{VB}(x|\mathcal{D}^N)) \right\rangle_{q(\mathcal{D}^N)}$$

for the marginal predictive distribution $p^{VB}(x|\mathcal{D}^N) = \langle p(x|w) \rangle_{\widehat{r}_w(w)}$. It follows from the log-sum inequality (17.9) that

$$\overline{GE}^{VB}(N) \leq \overline{GE}_{Joint}^{VB}(N). \quad (17.49)$$

Since the predictive distribution (17.42) is derived from the approximate posterior distribution $r^*(w; \tilde{w}^*)$ consisting of $p(\mathcal{H}|\mathcal{D}^N, \tilde{w}^*)$ instead of the minimizer $\widehat{r}_{\mathcal{H}}(\mathcal{H})$ of the free energy, it is conjectured that $\overline{GE}_{Joint}^{VB*}(N)$ provides a lower-bound to $\overline{GE}_{Joint}^{VB}(N)$. At least, the inequalities in Eq. (17.45) imply the affinity of the VB free energy and the generalization error measured by the KL divergence of the joint distributions. The generalization error of the marginal predictive distribution is generally upper-bounded by that of the joint predictive distribution as in Eq. (17.49). Although Eq. (17.48) shows that the average generalization error $\overline{GE}_{Joint}^{VB*}(N)$ of the approximate predictive distribution of VB learning with $r^*(w; \tilde{w}^*)$ is upper-bounded by that of Bayesian learning in the joint model, the relationship between $\overline{GE}_{Joint}^{VB*}(N)$ and $\overline{GE}_{Joint}^{VB}(N)$ is still unknown.

17.5 Relation to Other Analyses

In this section, we discuss the relationships of the asymptotic formulae in Sections 17.2 and 17.4 to the analyses of the Bayes free energy and the generalization error.

17.5.1 Asymptotic Analysis of Free Energy Bounds

Asymptotic upper-bounds of the Bayes free energy were obtained for some statistical models, including the GMM, HMM, and the Bayesian network (Yamazaki and Watanabe, 2003a,b, 2005). The upper-bounds are given by the following form:

$$\overline{F}^{Bayes}(N) \leq \lambda'_{Joint}^{Bayes} \log N + O(1), \quad (17.50)$$

where the coefficient $\lambda'^{\text{Bayes}}_{\text{Joint}}$ was identified for each model by analyzing the largest pole of the zeta function $\zeta_{\bar{E}}$ in Eq. (17.14) instead of ζ_E , by using the log-sum inequality (17.10) (Yamazaki and Watanabe, 2003a,b, 2005). Since the largest pole of $\zeta_{\bar{E}}$ provides a lower-bound for that of ζ_E , their analyses provided upper-bounds of $\bar{F}^{\text{Bayes}}(N)$ for the aforementioned models.

On the other hand, the asymptotic forms of the VB free energy were analyzed also for the same models as discussed in Chapters 15 and 16, each of which has the following form:

$$\bar{F}^{\text{VB}}(N) \leq \lambda'^{\text{VB}} \log N + O(1). \quad (17.51)$$

In most cases, asymptotic upper-bounds of $\bar{F}^{\text{Bayes}}(N)$ and $\bar{F}^{\text{VB}}(N)$ coincide, i.e., $\lambda'^{\text{Bayes}}_{\text{Joint}} = \lambda'^{\text{VB}}$ holds while Theorem 17.1 implies that $\lambda'^{\text{VB}} \leq \lambda'^{\text{Bayes}}_{\text{Joint}}$. Hence, it is suggested that this upper-bound is tight in some cases. The zeta function $\zeta_{\bar{E}}$ was also analyzed by the algebraic geometrical technique to evaluate the generalization error for estimating local latent variables (Yamazaki, 2016).

Moreover, the previous analyses of the VB free energy are based on the direct minimization of the free energy over the variational parameters (Chapters 14 through 16). Hence, the analyses are highly dependent on the concrete algorithm for the specific model and the choice of the prior distribution. In other words, it is required to parameterize the free energy explicitly by a finite number of variational parameters in such analyses. Analyzing the right-hand side of Eq. (17.24) is more general and is independent of the concrete algorithm for the specific model. It does not even require that the prior distribution $p(\mathbf{w})$ be conjugate since Theorem 17.1 holds for any prior. In such a case, the VB learning algorithm should be implemented with techniques for nonconjugacy such as the local variational approximation and the black-box variational inference (Section 2.1.7). In fact, for mixture models, the upper-bound in Theorem 15.10 averaged over the training samples can be obtained in more general cases. The mixture component $g(\mathbf{x}|\boldsymbol{\mu})$ can be generalized to any regular models, while in Chapter 15 it was generalized only to the exponential family.

17.5.2 Accuracy of Approximation

For several statistical models, tighter bounds or exact evaluations of the coefficient λ'^{Bayes} of the relative Bayes free energy in Eq. (17.5) have been obtained (Aoyagi and Watanabe, 2005; Yamazaki et al., 2010). If the relative Bayes free energy and VB free energy have the asymptotic forms, $\bar{F}^{\text{Bayes}}(N) = \lambda'^{\text{Bayes}} \log N + o(\log N)$ and $\bar{F}^{\text{VB}}(N) = \lambda'^{\text{VB}} \log N + o(\log N)$, respectively,

$\lambda'^{\text{Bayes}} \leq \lambda'^{\text{VB}}$ holds, and the approximation accuracy of VB learning to Bayesian learning can be evaluated by the gap between them:

$$\bar{F}^{\text{VB}}(N) - \bar{F}^{\text{Bayes}}(N) = (\lambda'^{\text{VB}} - \lambda'^{\text{Bayes}}) \log N + o(\log N).$$

From Eq. (17.17), this turns out to be the KL divergence from the approximate posterior to the true posterior. Such a comparison was first conducted for GMMs (Watanabe and Watanabe, 2004, 2006; Aoyagi and Nagata, 2012). A more detailed comparison was conducted for the Bernoulli mixture model discussed in Section 15.4 (Yamazaki and Kaji, 2013; Kaji et al., 2010). According to the authors' results, λ'^{VB} can be strictly greater than λ'^{Bayes} , while λ'^{VB} is not so large as $D/2$, where D is the number of parameters.⁵ The arguments in Section 17.2 imply that such a comparison can be extended to general latent variable models by examining the difference between $\min_{\bar{w}^* \in \mathcal{W}^*} \bar{F}_{\text{Joint}}^{\text{Bayes}}(N)$ and $\bar{F}^{\text{Bayes}}(N)$, which is related to the difference between $\lambda_{\text{Joint}}^{\text{Bayes}}$ and λ'^{Bayes} , i.e., the poles of ζ_E^- and ζ_E .

17.5.3 Average Generalization Error

Although the generalization performance of the VB learning was fully analyzed in the RRR model as discussed in Chapter 14, little has been known in other models. In Section 17.4, we derived an inequality that implies the relationship between the generalization error and the VB free energy for general latent variable models.

In the exact Bayesian learning, the universal relations (13.138) and (13.139) among the quartet, Bayes and Gibbs generalization losses and Bayes and Gibbs training losses, were proved as discussed in Section 13.5.5 (Watanabe, 2009). It is an important future work to explore such relationships among the quantities introduced in Section 17.4 for VB learning.

⁵ For the local latent variable model defined in Section 17.1, Corollary 17.2 combined with Eq. (13.125) implies that

$$2\lambda'^{\text{VB}} \leq D.$$

However, this is not true in general as we discussed for the RRR model in Chapter 14 (see Figure 14.8).

Appendix A James–Stein Estimator

The James–Stein (JS) estimator (James and Stein, 1961), a shrinkage estimator known to *dominate* the maximum likelihood (ML) estimator, has close relation to Bayesian learning. More specifically, it can be derived as an empirical Bayesian (EBayes) estimator (Efron and Morris, 1973).

Consider an M -dimensional Gaussian model with a Gaussian prior for the mean parameter:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \text{Gauss}_M(\mathbf{x}; \boldsymbol{\mu}, \sigma^2 \mathbf{I}_M) = (2\pi\sigma^2)^{-M/2} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|^2}{2\sigma^2}\right), \quad (\text{A.1})$$

$$p(\boldsymbol{\mu}|c^2) = \text{Gauss}_M(\boldsymbol{\mu}; \mathbf{0}, c^2 \mathbf{I}_M) = (2\pi c^2)^{-M/2} \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2c^2}\right), \quad (\text{A.2})$$

where the variance σ^2 of observation noise is assumed to be known. We perform empirical Bayesian learning to estimate the mean parameter $\boldsymbol{\mu}$ and the prior variance c^2 from observed samples $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. The joint distribution conditional to the hyperparameter is

$$\begin{aligned} p(\mathcal{D}, \boldsymbol{\mu}|c^2) &= p(\boldsymbol{\mu}|c^2) \prod_{n=1}^N p(\mathbf{x}^{(n)}|\boldsymbol{\mu}) \\ &= \frac{1}{(2\pi c^2)^{M/2} (2\pi\sigma^2)^{NM/2}} \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2c^2} - \sum_{n=1}^N \frac{\|\mathbf{x}^{(n)} - \boldsymbol{\mu}\|^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi c^2)^{M/2} (2\pi\sigma^2)^{NM/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)}\|^2 + \frac{N^2 \|\bar{\mathbf{x}}\|^2}{2\sigma^2(N + \sigma^2/c^2)}\right) \\ &\quad \cdot \exp\left(-\frac{N + \sigma^2/c^2}{2\sigma^2} \left\| \boldsymbol{\mu} - \frac{N\bar{\mathbf{x}}}{N + \sigma^2/c^2} \right\|^2\right), \end{aligned}$$

which implies that the posterior is Gaussian,

$$p(\boldsymbol{\mu}|\mathcal{D}, c^2) \propto p(\mathcal{D}, \boldsymbol{\mu}|c^2) \propto \exp\left(-\frac{N + \sigma^2/c^2}{2\sigma^2} \left\| \boldsymbol{\mu} - \frac{N\bar{\mathbf{x}}}{N + \sigma^2/c^2} \right\|^2\right),$$

with the mean given by

$$\widehat{\boldsymbol{\mu}} = \frac{N\bar{\mathbf{x}}}{N + \sigma^2/c^2} = \left(1 - \frac{\sigma^2}{Nc^2 + \sigma^2}\right)\bar{\mathbf{x}}. \quad (\text{A.3})$$

The marginal likelihood is computed as

$$\begin{aligned} p(\mathcal{D}|c^2) &= \int p(\mathcal{D}, \boldsymbol{\mu}|c^2) d\boldsymbol{\mu} \\ &= \frac{1}{(2\pi c^2)^{M/2} (2\pi \sigma^2)^{NM/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)}\|^2 + \frac{N^2 \|\bar{\mathbf{x}}\|^2}{2\sigma^2(N + \sigma^2/c^2)}\right) \\ &\quad \cdot \int \exp\left(-\frac{N + \sigma^2/c^2}{2\sigma^2} \left\|\boldsymbol{\mu} - \frac{N\bar{\mathbf{x}}}{N + \sigma^2/c^2}\right\|^2\right) d\boldsymbol{\mu} \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)}\|^2 + \frac{N^2 \|\bar{\mathbf{x}}\|^2}{2\sigma^2(N + \sigma^2/c^2)}\right)}{(2\pi c^2)^{M/2} (2\pi \sigma^2)^{(N-1)M/2} (N + \sigma^2/c^2)^{M/2}} \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2 - \frac{N\|\bar{\mathbf{x}}\|^2}{2\sigma^2} + \frac{N^2 \|\bar{\mathbf{x}}\|^2}{2\sigma^2(N + \sigma^2/c^2)}\right)}{(2\pi)^{M/2} (2\pi \sigma^2)^{(N-1)M/2} (Nc^2 + \sigma^2)^{M/2}} \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2 - \frac{N\sigma^2/c^2 \|\bar{\mathbf{x}}\|^2}{2\sigma^2(N + \sigma^2/c^2)}\right)}{(2\pi)^{M/2} (2\pi \sigma^2)^{(N-1)M/2} (Nc^2 + \sigma^2)^{M/2}} \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2 - \frac{N\|\bar{\mathbf{x}}\|^2}{2(Nc^2 + \sigma^2)}\right)}{(2\pi)^{M/2} (2\pi \sigma^2)^{(N-1)M/2} (Nc^2 + \sigma^2)^{M/2}} \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\|^2\right) \cdot \exp\left(-\frac{N\|\bar{\mathbf{x}}\|^2}{2(Nc^2 + \sigma^2)}\right)}{(2\pi \sigma^2)^{(N-1)M/2} \cdot (2\pi(Nc^2 + \sigma^2))^{M/2}}. \end{aligned} \quad (\text{A.4})$$

This implies that $v = \bar{\mathbf{x}}\sqrt{N/(Nc^2 + \sigma^2)}$ is a random variable subject to $\text{Gauss}_M(v; \mathbf{0}, \mathbf{I}_M)$. Since its (-2) nd order moment is equal to $\langle \|v\|^{-2} \rangle_{\text{Gauss}_M(v; \mathbf{0}, \mathbf{I}_M)} = (M-2)^{-1}$, we have

$$\left\langle \frac{Nc^2 + \sigma^2}{N \|\bar{\mathbf{x}}\|^2} \right\rangle_{p(\mathcal{D}|c^2)} = \frac{1}{M-2},$$

and therefore

$$\left\langle \frac{M-2}{N \|\bar{\mathbf{x}}\|^2} \right\rangle_{p(\mathcal{D}|c^2)} = \frac{1}{Nc^2 + \sigma^2}.$$

Accordingly, $(M-2)/N \|\bar{\mathbf{x}}\|^2$ is an unbiased estimator of the factor $(Nc^2 + \sigma^2)^{-1}$.

Replacing the factor $(Nc^2 + \sigma^2)^{-1}$ in Eq. (A.3) with its unbiased estimator $(M - 2)/N\|\bar{x}\|^2$, we obtain the JS estimator (with degree $M - 2$):

$$\hat{\mu}^{\text{JS}} = \left(1 - \frac{(M - 2)\sigma^2}{N\|\bar{x}\|^2}\right)\bar{x}. \quad (\text{A.5})$$

If we estimate c^2 by maximizing the marginal likelihood (A.4), we obtain the *positive-part JS estimator* (with degree M):

$$\hat{\mu}^{\text{PJS}} = \max\left(0, 1 - \frac{M\sigma^2}{N\|\bar{x}\|^2}\right)\bar{x}. \quad (\text{A.6})$$

The JS estimator has an interesting property. Let us first introduce terminology. Assume that we observed data \mathcal{D} generated from a distribution $p(\mathcal{D}|\boldsymbol{w})$ with unknown parameter \boldsymbol{w} . Consider two estimators $\hat{\boldsymbol{w}}_1 = \hat{\boldsymbol{w}}_1(\mathcal{D})$ and $\hat{\boldsymbol{w}}_2 = \hat{\boldsymbol{w}}_2(\mathcal{D})$, and measure some error criterion $E(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ from the true parameter value \boldsymbol{w}^* .

Definition A.1 (*Domination*) We say that the estimator $\hat{\boldsymbol{w}}_1$ dominates the other estimator $\hat{\boldsymbol{w}}_2$ if

$$\langle E(\hat{\boldsymbol{w}}_1(\mathcal{D}), \boldsymbol{w}^*) \rangle_{p(\mathcal{D}|\boldsymbol{w}^*)} \leq \langle E(\hat{\boldsymbol{w}}_2(\mathcal{D}), \boldsymbol{w}^*) \rangle_{p(\mathcal{D}|\boldsymbol{w}^*)} \quad \text{for arbitrary } \boldsymbol{w}^*, \\ \text{and} \quad \langle E(\hat{\boldsymbol{w}}_1(\mathcal{D}), \boldsymbol{w}^*) \rangle_{p(\mathcal{D}|\boldsymbol{w}^*)} < \langle E(\hat{\boldsymbol{w}}_2(\mathcal{D}), \boldsymbol{w}^*) \rangle_{p(\mathcal{D}|\boldsymbol{w}^*)} \quad \text{for a certain } \boldsymbol{w}^*.$$

Definition A.2 (*Efficiency*) We say that an estimator is efficient if no *unbiased* estimator dominates it.

Definition A.3 (*Admissibility*) We say that an estimator is admissible if no estimator dominates it.

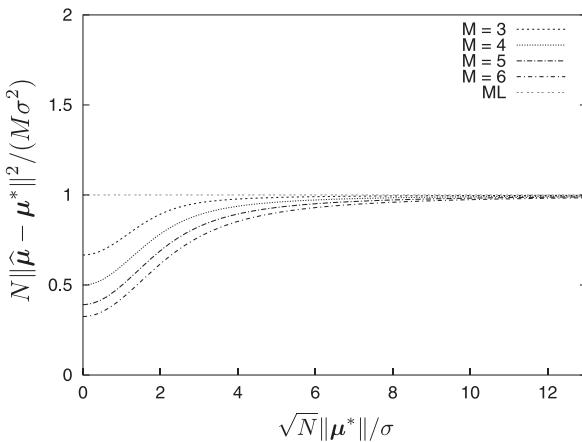


Figure A.1 Generalization error of James–Stein estimator.

Assume that $p(\mathcal{D}|\boldsymbol{w}) = \text{Gauss}_M(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2 \mathbf{I}_M)$. Then the ML estimator,

$$\hat{\boldsymbol{\mu}}^{\text{ML}} = \bar{\boldsymbol{x}}, \quad (\text{A.7})$$

is known to be *efficient* in terms of the mean squared error

$$E(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}^*) = \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|^2.$$

However, the ML estimator was proven to be inadmissible when $M \geq 3$, i.e., there exists at least one biased estimator that dominates the ML estimator (Stein, 1956). Subsequently, the JS estimator (A.5) was introduced as an estimator dominating the ML estimator (James and Stein, 1961).

Figure A.1 shows the normalized squared loss $N\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|^2/(M\sigma^2)$ of the ML estimator (A.7) and the JS estimator (A.5) as a function of a scaled true mean $\sqrt{N}\|\boldsymbol{\mu}^*\|/\sigma$. The ML estimator always gives error equal to one, while the JS estimator gives error dependent on the true value. We can see that the JS estimator dominates the ML estimator for $M \geq 3$. We can easily show that the positive-part JS estimator (A.6) dominates the JS estimator with the same degree.

Appendix B Metric in Parameter Space

In this appendix, we give a brief summary of the Kullback–Leibler (KL) divergence, the Fisher information, and the Jeffreys prior. The KL divergence is a common (pseudo-)distance measure between distributions, and the corresponding metric in the parameter space is given by the Fisher information. The Jeffreys prior—the uniform prior when the distance between distributions is measured by the KL divergence—is defined so as to reflect the nonuniformity of the density of the volume element in the parameter space.

B.1 Kullback–Leibler (KL) Divergence

The *KL divergence* between two distributions, $q(\mathbf{x})$ and $p(\mathbf{x})$, is defined as

$$\begin{aligned}\text{KL}(q(\mathbf{x})\|p(\mathbf{x})) &= \int q(\mathbf{x}) \log \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} \right) d\mathbf{x} \\ &= \int q(\mathbf{x}) \log \frac{1}{p(\mathbf{x})} d\mathbf{x} - \int q(\mathbf{x}) \log \frac{1}{q(\mathbf{x})} d\mathbf{x} \\ &\geq 0.\end{aligned}$$

If $q(\mathbf{x})$ is the true distribution, i.e., $\mathbf{x} \sim q(\mathbf{x})$, the first term is the average information gain for the one who has (possibly) wrong information (who believes $\mathbf{x} \sim p(\mathbf{x})$), and the second term is the average information gain, i.e., the *entropy*, for the one who has the correct information (who believes $\mathbf{x} \sim q(\mathbf{x})$). The KL divergence is not a proper distance metric, since it is not symmetric, i.e., for general $q(\mathbf{x})$ and $p(\mathbf{x})$,

$$\text{KL}(q(\mathbf{x})\|p(\mathbf{x})) \neq \text{KL}(p(\mathbf{x})\|q(\mathbf{x})).$$

B.2 Fisher Information

The *Fisher information* of a parametric distribution $p(\mathbf{x}|\mathbf{w})$ with its parameter $\mathbf{w} \in \mathbb{R}^D$ is defined as

$$\mathbb{S}_+^D \ni \mathbf{F} = \int \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \left(\frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \right)^\top p(\mathbf{x}|\mathbf{w}) d\mathbf{x}, \quad (\text{B.1})$$

where $\frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \in \mathbb{R}^D$ is the gradient (column) vector of $\log p(\mathbf{x}|\mathbf{w})$. Under the regularity conditions (see Section 13.4.1) on the statistical model $p(\mathbf{x}|\mathbf{w})$, the Fisher information can be written as

$$\mathbf{F} = - \int \frac{\partial^2 \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} p(\mathbf{x}|\mathbf{w}) d\mathbf{x}, \quad (\text{B.2})$$

where

$$\left(\frac{\partial^2 \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \right)_{ij} = \frac{\partial^2 \log p(\mathbf{x}|\mathbf{w})}{\partial w_i \partial w_j}.$$

This is because

$$\begin{aligned} & - \int \frac{\partial^2 \log p(\mathbf{x}|\mathbf{w})}{\partial w_i \partial w_j} p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \\ &= - \int \frac{\partial}{\partial w_j} \left(\frac{\frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial w_i}}{p(\mathbf{x}|\mathbf{w})} \right) p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \\ &= - \int \left(\frac{\frac{\partial^2 p(\mathbf{x}|\mathbf{w})}{\partial w_i \partial w_j}}{p(\mathbf{x}|\mathbf{w})} - \frac{\frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial w_i} \frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial w_j}}{p^2(\mathbf{x}|\mathbf{w})} \right) p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \\ &= - \int \frac{\partial^2 p(\mathbf{x}|\mathbf{w})}{\partial w_i \partial w_j} d\mathbf{x} + \int \frac{\frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial w_i} \frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial w_j}}{p^2(\mathbf{x}|\mathbf{w})} p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \\ &= - \frac{\partial^2}{\partial w_i \partial w_j} \int p(\mathbf{x}|\mathbf{w}) d\mathbf{x} + \int \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial w_i} \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial w_j} p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \\ &= \int \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial w_i} \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial w_j} p(\mathbf{x}|\mathbf{w}) d\mathbf{x}. \end{aligned}$$

B.3 Metric and Volume Element

For a small perturbation $\Delta \mathbf{w}$ of the parameter, the KL divergence between $p(\mathbf{x}|\mathbf{w})$ and $p(\mathbf{x}|\mathbf{w} + \Delta \mathbf{w})$ can be written as

$$\begin{aligned} \text{KL}(p(\mathbf{x}|\mathbf{w}) \| p(\mathbf{x}|\mathbf{w} + \Delta \mathbf{w})) &= \int p(\mathbf{x}|\mathbf{w}) \log \left(\frac{p(\mathbf{x}|\mathbf{w})}{p(\mathbf{x}|\mathbf{w} + \Delta \mathbf{w})} \right) d\mathbf{x} \\ &= \int p(\mathbf{x}|\mathbf{w}) \log \left(\frac{p(\mathbf{x}|\mathbf{w})}{p(\mathbf{x}|\mathbf{w}) + \frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}}^\top \Delta \mathbf{w} + \frac{1}{2} \Delta \mathbf{w}^\top \frac{\partial^2 p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}} \Delta \mathbf{w} + O(\|\Delta \mathbf{w}\|^3)} \right) d\mathbf{x} \\ &= - \int p(\mathbf{x}|\mathbf{w}) \log \left(1 + \frac{\frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}}^\top \Delta \mathbf{w}}{p(\mathbf{x}|\mathbf{w})} + \frac{1}{2} \Delta \mathbf{w}^\top \frac{\frac{\partial^2 p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}}}{p(\mathbf{x}|\mathbf{w})} \Delta \mathbf{w} + O(\|\Delta \mathbf{w}\|^3) \right) d\mathbf{x} \\ &= - \int p(\mathbf{x}|\mathbf{w}) \left(\frac{\frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}}^\top}{p(\mathbf{x}|\mathbf{w})} \Delta \mathbf{w} + \frac{1}{2} \Delta \mathbf{w}^\top \left(\frac{\frac{\partial^2 p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}}}{p(\mathbf{x}|\mathbf{w})} - \frac{\frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \frac{\partial p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}}^\top}{p^2(\mathbf{x}|\mathbf{w})} \right) \Delta \mathbf{w} \right) d\mathbf{x} \\ &\quad + O(\|\Delta \mathbf{w}\|^3) \end{aligned}$$

$$\begin{aligned}
&= - \left(\frac{\partial}{\partial \mathbf{w}} \int p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \right)^\top \Delta \mathbf{w} - \frac{1}{2} \Delta \mathbf{w}^\top \left(\frac{\partial^2}{\partial \mathbf{w} \partial \mathbf{w}} \int p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \right) \Delta \mathbf{w} \\
&\quad + \frac{1}{2} \Delta \mathbf{w}^\top \left(\int \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}} \frac{\partial \log p(\mathbf{x}|\mathbf{w})}{\partial \mathbf{w}}^\top p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \right) \Delta \mathbf{w} + O(\|\Delta \mathbf{w}\|^3) \\
&= \frac{1}{2} \Delta \mathbf{w}^\top \mathbf{F} \Delta \mathbf{w} + O(\|\Delta \mathbf{w}\|^3).
\end{aligned}$$

Therefore, the Fisher information corresponds to the *metric* of the space of distributions when the distance is measured by the KL divergence (Jeffreys, 1946).

When we adopt the Fisher information as the metric, the *volume element* for integrating functions is given by

$$dV = \frac{1}{\sqrt{2}} \sqrt{\det(\mathbf{F})} d\mathbf{w}, \quad (\text{B.3})$$

where $\frac{1}{\sqrt{2}} \sqrt{\det(\mathbf{F})}$ corresponds to the *density*.

B.4 Jeffreys Prior

The prior,

$$p(\mathbf{w}) \propto \sqrt{\det(\mathbf{F})}, \quad (\text{B.4})$$

proportional to the density of the volume element (B.3), is called the *Jeffreys prior* (Jeffreys, 1946). The Jeffreys prior assigns the equal probability to the unit volume element at any point in the parameter space, i.e., it is the *uniform prior* in the distribution space. Since the uniformity is defined not in the parameter space but in the distribution space, the Jeffreys prior is invariant under parameter transformation. Accordingly, the Jeffreys prior is said to be the parameterization invariant *noninformative prior*.

For *singular models*, the Fisher information can have zero eigenvalues, which makes the Jeffreys prior zero. In some models, including the matrix factorization model, zero eigenvalues appear everywhere in the parameter space (see Example B.2). In such cases, we ignore the *common* zero eigenvalues and redefine the (generalized) Jeffrey prior by

$$p(\mathbf{w}) \propto \sqrt{\prod_{d=1}^{\bar{D}} \lambda_d}, \quad (\text{B.5})$$

where λ_d is the d th largest eigenvalue of the Fisher information \mathbf{F} , and \bar{D} is the maximum number of positive eigenvalues over the whole parameter space.

Example B.1 (Jeffreys prior for one-dimensional Gaussian distribution) The Fisher information of the Gaussian distribution,

$$p(x|\mu, \sigma^2) = \text{Gauss}_1(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

is calculated as follows. The derivatives of the log likelihood are

$$\begin{aligned}
\frac{\partial \log p(x|\mu, \sigma^2)}{\partial \mu} &= \frac{\partial}{\partial \mu} \left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \\
&= \frac{x-\mu}{\sigma^2},
\end{aligned}$$

$$\begin{aligned}\frac{\partial \log p(x|\mu, \sigma^2)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2} \left(-\frac{1}{2} \log \sigma^2 - \frac{(x-\mu)^2}{2\sigma^2} \right) \\ &= -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}, \\ \frac{\partial^2 \log p(x|\mu, \sigma^2)}{\partial \mu^2} &= -\frac{1}{\sigma^2}, \\ \frac{\partial^2 \log p(x|\mu, \sigma^2)}{\partial \mu \partial \sigma^2} &= -\frac{x-\mu}{\sigma^4}, \\ \frac{\partial^2 \log p(x|\mu, \sigma^2)}{\partial (\sigma^2)^2} &= \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6},\end{aligned}$$

and therefore

$$\begin{aligned}\mathbf{F} &= \left\langle \begin{pmatrix} \frac{1}{\sigma^2} & \frac{x-\mu}{\sigma^4} \\ \frac{x-\mu}{\sigma^4} & \frac{(x-\mu)^2}{\sigma^6} - \frac{1}{2\sigma^4} \end{pmatrix} \right\rangle_{p(x|\mu, \sigma^2)} \\ &= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^4} - \frac{1}{2\sigma^4} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.\end{aligned}$$

Thus, the Jeffreys priors for $p(x|\mu)$, $p(x|\sigma^2)$, and $p(\mu|\sigma^2)$ are

$$\begin{aligned}p(\mu) &\propto \sqrt{\mathbf{F}_{\mu, \mu}} \propto 1, \\ p(\sigma^2) &\propto \sqrt{\mathbf{F}_{\sigma^2, \sigma^2}} \propto \frac{1}{\sigma^2}, \\ p(\mu, \sigma^2) &\propto \sqrt{\det(\mathbf{F})} \propto \frac{1}{\sigma^3},\end{aligned}$$

respectively.

Example B.2 (Jeffreys prior for one-dimensional matrix factorization model) The Fisher information of the one-dimensional matrix factorization (MF) model,

$$p(V|A, B) = \text{Gauss}_1(V; BA, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(V-BA)^2}{2\sigma^2}\right), \quad (\text{B.6})$$

is calculated as follows. The derivatives of the log likelihood are

$$\begin{aligned}\frac{\partial \log p(V|A, B)}{\partial A} &= \sigma^{-2}(V-BA)B, \\ \frac{\partial \log p(V|A, B)}{\partial B} &= \sigma^{-2}(V-BA)A,\end{aligned}$$

and therefore

$$\begin{aligned}\mathbf{F} &= \frac{1}{\sigma^4} \left\langle \begin{pmatrix} (V-BA)^2 B^2 & (V-BA)^2 BA \\ (V-BA)^2 BA & (V-BA)^2 A^2 \end{pmatrix} \right\rangle_{\text{Gauss}_1(V; BA, \sigma^2)} \\ &= \frac{1}{\sigma^2} \begin{pmatrix} B^2 & BA \\ BA & A^2 \end{pmatrix}.\end{aligned}$$

The Fisher information \mathbf{F} has eigenvalues $\lambda_1 = \sigma^{-2}(A^2 + B^2)$ and $\lambda_2 = 0$, since

$$\begin{aligned}\det(\sigma^2 \mathbf{F} - \lambda \mathbf{I}_2) &= \det \begin{pmatrix} B^2 - \lambda & BA \\ BA & A^2 - \lambda \end{pmatrix} = (B^2 - \lambda)(A^2 - \lambda) - B^2 A^2 \\ &= \lambda^2 - (A^2 + B^2)\lambda \\ &= (\lambda - (A^2 + B^2))\lambda.\end{aligned}$$

The common (over the whole parameter space) zero eigenvalue comes from the invariance of the MF model under the transformation $(A, B) \rightarrow (sA, s^{-1}B)$ for any $s \neq 0$. By adopting the generalized definition (B.5) of the Jeffreys prior, the distribution proportional to

$$p(A, B) \propto \sqrt{A^2 + B^2} \quad (\text{B.7})$$

is the parameterization invariant noninformative prior.

The Jeffreys prior is often improper, i.e., the integral of the unnormalized prior over the parameter domain diverges, and therefore the normalization factor cannot be computed, as in Examples B.1 and B.2.

Appendix C Detailed Description of Overlap Method

Let $\mathbf{V} \in \mathbb{R}^{L \times M}$ be the observed matrix, where L and M correspond to the dimensionality D of the observation space and the number N of samples as follows:

$$\begin{aligned} L = D, M = N & \quad \text{if} \quad D \leq N, \\ L = N, M = D & \quad \text{if} \quad D > N. \end{aligned} \tag{C.1}$$

Let

$$\mathbf{V} = \sum_{h=1}^L \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top \tag{C.2}$$

be the singular value decomposition (SVD) of \mathbf{V} . The overlap (OL) method (Hoyle, 2008) computes the following approximation to the negative logarithm of the marginal likelihood (8.90) over the hypothetical model rank $H = 1, \dots, L$:¹

$$\begin{aligned} 2F^{\text{OL}}(H) & \approx -2 \log p(\mathbf{V}) \\ & = (LM - H(L - H - 2)) \log(2\pi) + L \log \pi - 2 \sum_{h=1}^H \log \left(\frac{\Gamma((M-h+1)/2)}{\Gamma(M-L-h+1)/2} \right) \\ & \quad + H(M - L)(1 - \log(M - L)) + \sum_{h=1}^H \sum_{l=H+1}^L \log(\gamma_h^2 - \gamma_l^2) \\ & \quad + (M - L) \sum_{h=1}^H \log \gamma_h^2 + (M - H) \sum_{h=1}^H \log \left(\frac{1}{\hat{\sigma}^2 \text{OL}} - \frac{1}{\lambda_h^{\text{OL}}} \right) \\ & \quad - \sum_{h=1}^H \left(\frac{1}{\hat{\sigma}^2 \text{OL}} - \frac{1}{\lambda_h^{\text{OL}}} \right) \gamma_h^2 + (L + 2) \left(\sum_{h=1}^H \log \hat{\lambda}_h^{\text{OL}} + (M - H) \log \hat{\sigma}^2 \text{OL} \right) \\ & \quad + \sum_{l=1}^L \frac{\gamma_l^2}{\hat{\sigma}^2 \text{OL}}, \end{aligned} \tag{C.3}$$

where $\Gamma(\cdot)$ denotes the *Gamma function*, and $\{\hat{\lambda}_h^{\text{OL}}\}_{h=1}^H$ and $\hat{\sigma}^2 \text{OL}$ are estimators for $\{\lambda_h = b_h^2 + \sigma^2\}_{h=1}^H$ and σ^2 , respectively, computed by iterating the following updates until convergence:

¹ Our description is slightly different from Hoyle (2008), because the MF model (6.1) does not have the mean parameter shared over the samples.

Algorithm 23 Overlap method.

-
- 1: Prepare the observed matrix $\mathbf{V} \in \mathbb{R}^{L \times M}$, following the rule (C.1).
 - 2: Compute the SVD (C.2) of \mathbf{V} .
 - 3: Compute $F^{\text{OL}}(0)$ by Eq. (C.6).
 - 4: **for** $H = 1$ to L **do**
 - 5: Initialize the noise variance to $\widehat{\sigma}^2 \text{ OL} = 10^{-4} \cdot \sum_{h=1}^L \gamma_h^2 / (LM)$.
 - 6: Iterate Eq. (C.4) for $h = 1, \dots, H$, and Eq. (C.5) until convergence or any $\widehat{\lambda}_h^{\text{OL}}$ becomes a complex number.
 - 7: Compute $F^{\text{OL}}(H)$ by Eq. (C.3) if all $\{\widehat{\lambda}_h^{\text{OL}}\}_{h=1}^H$ are real numbers. Otherwise, set $F^{\text{OL}}(H) = \infty$.
 - 8: **end for**
 - 9: Estimate the rank by $\widehat{H}^{\text{OL}} = \min_{H \in \{0, \dots, L\}} F^{\text{OL}}(H)$.
-

$$\begin{aligned}\widehat{\lambda}_h^{\text{OL}} &= \frac{\gamma_h^2}{2(L+2)} \left(1 - \frac{(M-H-(L+2))\widehat{\sigma}^2 \text{ OL}}{\gamma_h^2} \right. \\ &\quad \left. + \sqrt{\left(1 - \frac{(M-H-(L+2))\widehat{\sigma}^2 \text{ OL}}{\gamma_h^2} \right)^2 - \frac{4(L+2)\widehat{\sigma}^2 \text{ OL}}{\gamma_h^2}} \right),\end{aligned}\quad (\text{C.4})$$

$$\widehat{\sigma}^2 \text{ OL} = \frac{1}{(M-H)} \left(\sum_{l=1}^L \frac{\gamma_l^2}{L} - \sum_{h=1}^H \widehat{\lambda}_h^{\text{OL}} \right). \quad (\text{C.5})$$

When iterating Eqs. (C.4) and (C.5), $\widehat{\lambda}_h^{\text{OL}}$ can become a complex number. In such a case, the hypothetical H is rejected. Otherwise, Eq. (C.3) is evaluated after convergence. For the null hypothesis, i.e., $H = 0$, the negative log likelihood is given by

$$2F^{\text{OL}}(0) = -2 \log p(\mathbf{V}) = LM \left(\log \left(\frac{2\pi}{LM} \sum_{l=1}^L \gamma_l^2 \right) + 1 \right). \quad (\text{C.6})$$

The estimated rank \widehat{H}^{OL} is the minimizer of $F^{\text{OL}}(H)$ over $H = 0, \dots, L$.

Algorithm 23 summarizes the procedure.

Appendix D Optimality of Bayesian Learning

Bayesian learning is deduced from the basic probability theory, and therefore it is optimal in terms of generalization performance under the assumption that the model and the prior are set reasonably.

Consider a distribution of problems where the true distribution is written as $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{w}^*)$ with the true parameter \mathbf{w}^* subject to $q(\mathbf{w}^*)$. Although we usually omit the dependency description on the true distribution or the true parameter, the average generalization error, Eq. (13.13), naturally depends on the true distribution, so we here denote the dependence explicitly as $\overline{\text{GE}}(N; \mathbf{w}^*)$. Let

$$\overline{\overline{\text{GE}}}(N) = \left\langle \overline{\text{GE}}(N; \mathbf{w}^*) \right\rangle_{q(\mathbf{w}^*)} \quad (\text{D.1})$$

be the average of the average generalization error over the distribution $q(\mathbf{w}^*)$ of the true parameter.

Theorem D.1 *If we know the distribution $q(\mathbf{w}^*)$ of the true parameter and use it as the prior distribution, i.e., $p(\mathbf{w}) = q(\mathbf{w})$, then Bayesian learning minimizes the average generalization error (D.1) over $q(\mathbf{w}^*)$, i.e.,*

$$\overline{\overline{\text{GE}}}^{\text{Bayes}}(N) \leq \overline{\overline{\text{GE}}}^{\text{Other}}(N), \quad (\text{D.2})$$

where $\overline{\overline{\text{GE}}}^{\text{Other}}(N)$ denotes the average generalization error of any (other) learning algorithm.

Proof Let $\mathbf{X}^N = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$ be the N training samples. Regarding the new test sample as the $(N + 1)$ th sample, we can write the Bayes predictive distribution as follows:

$$\begin{aligned} p^{\text{Bayes}}(\mathbf{x}^{(N+1)} | \mathbf{X}^N) &= \int p(\mathbf{x}^{(N+1)} | \mathbf{w}) p(\mathbf{w} | \mathbf{X}^N) d\mathbf{w} \\ &= \frac{\int p(\mathbf{w}) \prod_{n=1}^{N+1} p(\mathbf{x}^{(n)} | \mathbf{w}) d\mathbf{w}}{\int p(\mathbf{w}') \prod_{n=1}^N p(\mathbf{x}^{(n)} | \mathbf{w}') d\mathbf{w}'} \\ &= \frac{p(\mathbf{X}^{N+1})}{p(\mathbf{X}^N)}, \end{aligned} \quad (\text{D.3})$$

where $p(\mathbf{X}^N) = \int p(\mathbf{w})p(\mathbf{X}^N|\mathbf{w})d\mathbf{w}$ is the marginal likelihood. The average generalization error (D.1) of a learning algorithm with its predictive distribution $r(\mathbf{x})$ is given by

$$\begin{aligned}\overline{\overline{\text{GE}}}(N) &= \left\langle \log \frac{p(\mathbf{x}^{(N+1)}|\mathbf{w}^*)}{r(\mathbf{x}^{(N+1)})} \right\rangle_{p(\mathbf{X}^{N+1}|\mathbf{w}^*)q(\mathbf{w}^*)} \\ &= - \int \left(\int p(\mathbf{X}^{N+1}|\mathbf{w}^*)q(\mathbf{w}^*)d\mathbf{w}^* \right) \log r(\mathbf{x}^{(N+1)})d\mathbf{X}^{N+1} - (N+1)S, \\ &= - \int q(\mathbf{X}^{N+1}) \log r(\mathbf{x}^{(N+1)})d\mathbf{X}^{N+1} - (N+1)S,\end{aligned}\quad (\text{D.4})$$

where

$$q(\mathbf{X}^N) = \int p(\mathbf{X}^N|\mathbf{w}^*)q(\mathbf{w}^*)d\mathbf{w}^*$$

is the marginal likelihood with the *true* prior distribution $q(\mathbf{w}^*)$, and

$$S = - \langle \log p(\mathbf{x}|\mathbf{w}^*) \rangle_{p(\mathbf{x}|\mathbf{w}^*)q(\mathbf{w}^*)}$$

is the entropy, which does not depend on the predictive distribution $r(\mathbf{x})$. Eq. (D.4) can be written as

$$\begin{aligned}\overline{\overline{\text{GE}}}(N) &= - \int q(\mathbf{X}^N) \frac{q(\mathbf{X}^{N+1})}{q(\mathbf{X}^N)} \log r(\mathbf{x}^{(N+1)})d\mathbf{X}^{N+1} - (N+1)S \\ &= - \left\langle \int \frac{q(\mathbf{X}^{N+1})}{q(\mathbf{X}^N)} \log r(\mathbf{x}^{(N+1)})d\mathbf{x}^{(N+1)} \right\rangle_{q(\mathbf{X}^N)} - (N+1)S \\ &= \left\langle \int \frac{q(\mathbf{X}^{N+1})}{q(\mathbf{X}^N)} \log \frac{\frac{q(\mathbf{X}^{N+1})}{q(\mathbf{X}^N)}}{r(\mathbf{x}^{(N+1)})} d\mathbf{x}^{(N+1)} \right\rangle_{q(\mathbf{X}^N)} + \text{const.}\end{aligned}\quad (\text{D.5})$$

Since the first term is the KL divergence between $q(\mathbf{X}^{N+1})/q(\mathbf{X}^N)$ and $r(\mathbf{x}^{(N+1)})$, Eq. (D.5) is minimized when

$$r(\mathbf{x}^{(N+1)}) = \frac{q(\mathbf{X}^{N+1})}{q(\mathbf{X}^N)} = q^{\text{Bayes}}(\mathbf{x}^{(N+1)}|\mathbf{X}^N),\quad (\text{D.6})$$

where $q^{\text{Bayes}}(\mathbf{x}^{(N+1)}|\mathbf{X}^N)$ is the Bayes predictive distribution (D.3) with the prior distribution set to the distribution of the true parameter, i.e., $p(\mathbf{w}) = q(\mathbf{w})$. Thus, we have proved that no other learning method can give better generalization error than Bayesian learning with the true prior. \square

A remark is that, since we usually do not know the true distribution and the true prior (the distribution of the true parameter), it is not surprising that an approximation method, e.g., variational Bayesian learning, to Bayesian learning provides better generalization performance than Bayesian learning with a nontrue prior in some situations.

Bibliography

- Akaho, S., and Kappen, H. J. 2000. Nonmonotonic Generalization Bias of Gaussian Mixture Models. *Neural Computation*, **12**, 1411–1427.
- Akaike, H. 1974. A New Look at Statistical Model. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.
- Akaike, H. 1980. Likelihood and Bayes Procedure. Pages 143–166 of: Bernald, J. M. (ed.), *Bayesian Statistics*. Valencia, Italy: University Press.
- Alzer, H. 1997. On Some Inequalities for the Gamma and Psi Functions. *Mathematics of Computation*, **66**(217), 373–389.
- Amari, S., Park, H., and Ozeki, T. 2002. Geometrical Singularities in the Neuromanifold of Multilayer Perceptrons. Pages 343–350 of: *Advances in NIPS*, vol. 14. Cambridge, MA: MIT Press.
- Aoyagi, M., and Nagata, K. 2012. Learning Coefficient of Generalization Error in Bayesian Estimation and Vandermonde Matrix-Type Singularity. *Neural Computation*, **24**(6), 1569–1610.
- Aoyagi, M., and Watanabe, S. 2005. Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation. *Neural Networks*, **18**(7), 924–933.
- Asuncion, A., and Newman, D.J. 2007. *UCI Machine Learning Repository*. www.ics.uci.edu/~mlearn/MLRepository.html
- Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. 2009. On Smoothing and Inference for Topic Models. Pages 27–34 of: *Proceedings of UAI*. Stockholm, Sweden: Morgan Kaufmann Publishers Inc.
- Attias, H. 1999. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. Pages 21–30 of: *Proceedings of UAI*. Stockholm, Sweden: Morgan Kaufmann Publishers Inc.
- Babacan, S. D., Nakajima, S., and Do, M. N. 2012a. Probabilistic Low-Rank Subspace Clustering. Pages 2753–2761 of: *Advances in Neural Information Processing Systems 25*. Lake Tahoe, NV: NIPS Foundation.
- Babacan, S. D., Luessi, M., Molina, R., and Katsaggelos, A. K. 2012b. Sparse Bayesian Methods for Low-Rank Matrix Estimation. *IEEE Transactions on Signal Processing*, **60**(8), 3964–3977.
- Baik, J., and Silverstein, J. W. 2006. Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models. *Journal of Multivariate Analysis*, **97**(6), 1382–1408.

- Baldi, P. F., and Hornik, K. 1995. Learning in Linear Neural Networks: A Survey. *IEEE Transactions on Neural Networks*, **6**(4), 837–858.
- Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. 2005. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, **6**, 1705–1749.
- Beal, M. J. 2003. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London.
- Bicego, M., Lovato, P., Ferrarini, A., and Delledonne, M. 2010. Biclustering of Expression Microarray Data with Topic Models. Pages 2728–2731 of: *Proceedings of ICPR*. Istanbul, Turkey: ICPR.
- Bickel, P., and Chernoff, H. 1993. *Asymptotic Distribution of the Likelihood Ratio Statistic in a Prototypical Non Regular Problem*. New Delhi, India: Wiley Eastern Limited.
- Bishop, C. M. 1999a. Bayesian Principal Components. Pages 382–388 of: *Advances in NIPS*, vol. 11. Denver, CO: NIPS Foundation.
- Bishop, C. M. 1999b. Variational Principal Components. Pages 514–509 of: *Proceedings of International Conference on Artificial Neural Networks*, vol. 1. Edinburgh, UK: Computing and Control Engineering Journal.
- Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.
- Bishop, C. M., and Tipping, M. E. 2000. Variational Relevance Vector Machines. Pages 46–53 of: *Proceedings of the Sixteenth Conference Annual Conference on Uncertainty in Artificial Intelligence*. Stanford, CA: Morgan Kaufmann Publishers Inc.
- Blei, D. M., and Jordan, M. I. 2005. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, **1**, 121–144.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.
- Bouchaud, J. P., and Potters, M. 2003. *Theory of Financial Risk and Derivative Pricing—From Statistical Physics to Risk Management*, 2nd edn. Cambridge, UK: University Press.
- Brown, L. D. 1986. *Fundamentals of Statistical Exponential Families*. IMS Lecture Notes–Monograph Series 9. Beachwood, OH: Institute of Mathematical Statistics.
- Candès, E. J., Li, X., Ma, Y., and Wright, J. 2011. Robust Principal Component Analysis? *Journal of the ACM*, **58**(3), 1–37.
- Carroll, J. D., and Chang, J. J. 1970. Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization of “Eckart–Young” Decomposition. *Psychometrika*, **35**, 283–319.
- Chen, X., Hu, X., Shen, X., and Rosen, G. 2010. Probabilistic Topic Modeling for Genomic Data Interpretation. Pages 149–152 of: *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.
- Chib, S. 1995. Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**(432), 1313–1321.
- Chu, W., and Ghahramani, Z. 2009. Probabilistic Models for Incomplete Multi-dimensional Arrays. Pages 89–96. In: *Proceedings of International Conference on Artificial Intelligence and Statistics*. Clearwater Beach, FL: Proceedings of Machine Learning Research.

- Courant, R., and Hilbert, D. 1953. *Methods of Mathematical Physics, Volume 1*. New York: Wiley.
- Cramer, H. 1949. *Mathematical Methods of Statistics*. Princeton, NJ: University Press.
- Dacunha-Castelle, D., and Gassiat, E. 1997. Testing in Locally Conic Models, and Application to Mixture Models. *Probability and Statistics*, **1**, 285–317.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum Likelihood for Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **39-B**, 1–38.
- Dharmadhikari, S., and Joag-Dev, K. 1988. *Unimodality, Convexity, and Applications*. Cambridge, MA: Academic Press.
- Ding, X., He, L., and Carin, L. 2011. Bayesian Robust Principal Component Analysis. *IEEE Transactions on Image Processing*, **20**(12), 3419–3430.
- Drexler, F. J. 1978. A Homotopy Method for the Calculation of All Zeros of Zero-Dimensional Polynomial Ideals. Pages 69–93 of: Wacker, H. J. (ed.), *Continuation Methods*. New York: Academic Press.
- D’Souza, A., Vijayakumar, S., and Schaal, S. 2004. The Bayesian Backfitting Relevance Vector Machine. In: *Proceedings of the 21st International Conference on Machine Learning*. Banff, AB: Association for Computing Machinery.
- Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.
- Efron, B., and Morris, C. 1973. Stein’s Estimation Rule and its Competitors—An Empirical Bayes Approach. *Journal of the American Statistical Association*, **68**, 117–130.
- Elhamifar, E., and Vidal, R. 2013. Sparse Subspace Clustering: Algorithm, Theory, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(11), 2765–2781.
- Felzenszwalb, P. F., and Huttenlocher, D. P. 2004. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, **59**(2), 167–181.
- Fukumizu, K. 1999. Generalization Error of Linear Neural Networks in Unidentifiable Cases. Pages 51–62 of: *Proceedings of International Conference on Algorithmic Learning Theory*. Tokyo, Japan: Springer.
- Fukumizu, K. 2003. Likelihood Ratio of Unidentifiable Models and Multilayer Neural Networks. *Annals of Statistics*, **31**(3), 833–851.
- Garcia, C. B., and Zangwill, W. I. 1979. Determining All Solutions to Certain Systems of Nonlinear Equations. *Mathematics of Operations Research*, **4**, 1–14.
- Gershman, S. J., and Blei, D. M. 2012. A Tutorial on Bayesian Nonparametric Models. *Journal of Mathematical Psychology*, **56**(1), 1–12.
- Ghahramani, Z., and Beal, M. J. 2001. Graphical Models and Variational Methods. Pages 161–177 of: *Advanced Mean Field Methods*. Cambridge, MA: MIT Press.
- Girolami, M. 2001. A Variational Method for Learning Sparse and Overcomplete Representations. *Neural Computation*, **13**(11), 2517–2532.
- Girolami, M., and Kaban, A. 2003. On an Equivalence between PLSI and LDA. Pages 433–434 of: *Proceedings of SIGIR*, New York and Toronto, ON: Association for Computing Machinery.

- Gopalan, P., Hofman, J. M., and Blei, D. M. 2013. Scalable Recommendation with Poisson Factorization. *arXiv:1311.1704 [cs.IR]*.
- Griffiths, T. L., and Steyvers, M. 2004. Finding Scientific Topics. *PNAS*, **101**, 5228–5235.
- Gunji, T., Kim, S., Kojima, M., Takeda, A., Fujisawa, K., and Mizutani, T. 2004. PHoM—A Polyhedral Homotopy Continuation Method. *Computing*, **73**, 57–77.
- Gupta, A. K., and Nagar, D. K. 1999. *Matrix Variate Distributions*. London, UK: Chapman and Hall/CRC.
- Hagiwara, K. 2002. On the Problem in Model Selection of Neural Network Regression in Overrealizable Scenario. *Neural Computation*, **14**, 1979–2002.
- Hagiwara, K., and Fukumizu, K. 2008. Relation between Weight Size and Degree of Over-Fitting in Neural Network Regression. *Neural Networks*, **21**(1), 48–58.
- Han, T. S., and Kobayashi, K. 2007. *Mathematics of Information and Coding*. Providence, RI: American Mathematical Society.
- Harshman, R. A. 1970. Foundations of the PARAFAC Procedure: Models and Conditions for an “Explanatory” Multimodal Factor Analysis. *UCLA Working Papers in Phonetics*, **16**, 1–84.
- Hartigan, J. A. 1985. A Failure of Likelihood Ratio Asymptotics for Normal Mixtures. Pages 807–810 of: *Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer*. Berkeley, CA: Springer.
- Hastie, T., and Tibshirani, R. 1986. Generalized Additive Models. *Statistical Science*, **1**(3), 297–318.
- Hinton, G. E., and van Camp, D. 1993. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. Pages 5–13 of: *Proceedings of COLT*. Santa Cruz, CA.
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. 2013. Stochastic Variational Inference. *Journal of Machine Learning Research*, **14**, 1303–1347.
- Hofmann, T. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, **42**, 177–196.
- Hosino, T., Watanabe, K., and Watanabe, S. 2005. Stochastic Complexity of Variational Bayesian Hidden Markov Models. In: *Proceedings of IJCNN*. Montreal, QC.
- Hosino, T., Watanabe, K., and Watanabe, S. 2006a. Free Energy of Stochastic Context Free Grammar on Variational Bayes. Pages 407–416 of: *Proceedings of ICONIP*. Hong Kong, China: Springer.
- Hosino, T., Watanabe, K., and Watanabe, S. 2006b. Stochastic Complexity of Hidden Markov Models on the Variational Bayesian Learning (in Japanese). *IEICE Transactions on Information and Systems*, **J89-D**(6), 1279–1287.
- Hotelling, H. 1933. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, **24**, 417–441.
- Hoyle, D. C. 2008. Automatic PCA Dimension Selection for High Dimensional Data and Small Sample Sizes. *Journal of Machine Learning Research*, **9**, 2733–2759.
- Hoyle, D. C., and Rattray, M. 2004. Principal-Component-Analysis Eigenvalue Spectra from Data with Symmetry-Breaking Structure. *Physical Review E*, **69**(026124).
- Huynh, T., Mario, F., and Schiele, B. 2008. Discovery of Activity Patterns Using Topic Models. Pages 9–10. In: *International Conference on Ubiquitous Computing (Ubicomp)*. New York and Seoul, South Korea: Association for Computer Machinery.

- Hyvärinen, A., Karhunen, J., and Oja, E. 2001. *Independent Component Analysis*. New York: Wiley.
- Ibragimov, I. A. 1956. On the Composition of Unimodal Distributions. *Theory of Probability and Its Applications*, **1**(2), 255–260.
- Ilin, A., and Raiko, T. 2010. Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *Journal of Machine Learning Research*, **11**, 1957–2000.
- Ito, H., Amari, S., and Kobayashi, K. 1992. Identifiability of Hidden Markov Information Sources and Their Minimum Degrees of Freedom. *IEEE Transactions on Information Theory*, **38**(2), 324–333.
- Jaakkola, T. S., and Jordan, M. I. 2000. Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing*, **10**, 25–37.
- James, W., and Stein, C. 1961. Estimation with Quadratic Loss. Pages 361–379 of: *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley: University of California Press.
- Jeffreys, H. 1946. An Invariant Form for the Prior Probability in Estimation Problems. Pages 453–461 of: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186. London, UK: Royal Society.
- Jensen, F. V. 2001. *Bayesian Networks and Decision Graphs*. Springer.
- Johnstone, I. M. 2001. On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *Annals of Statistics*, **29**, 295–327.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. 1999. Introduction to Variational Methods for Graphical Models. *Machine Learning*, **37**, 183–233.
- Kaji, D., Watanabe, K., and Watanabe, S. 2010. Phase Transition of Variational Bayes Learning in Bernoulli Mixture. *Australian Journal of Intelligent Information Processing Systems*, **11**(4), 35–40.
- Khan, M. E., Babanezhad, R., Lin, W., Schmidt, M., and Sugiyama, M. 2016. Faster Stochastic Variational Inference Using Proximal-Gradient Methods with General Divergence Functions. Pages 309–318. In: *Proceedings of UAI*. New York: AUAI Press.
- Kim, Y. D., and Choi, S. 2014. Scalable Variational Bayesian Matrix Factorization with Side Information. Pages 493–502 of: *Proceedings of AISTATS*. Reykjavik, Iceland: Proceedings of Machine Learning Research.
- Kingma, D. P., and Welling, M. 2014. Auto-Encoding Variational Bayes. In: *International Conference on Learning Representations (ICLR)*. arXiv:1412.6980
- Kolda, T. G., and Bader, B. W. 2009. Tensor Decompositions and Applications. *SIAM Review*, **51**(3), 455–500.
- Krestel, R., Fankhauser, P., and Nejdl, W. 2009. Latent Dirichlet Allocation for Tag Recommendation. Pages 61–68 of: *Proceedings of the Third ACM Conference on Recommender Systems*. New York: Association for Computing Machinery.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. Pages 1097–1105 of: *Advances in NIPS*. Lake Tahoe, NV: NIPS Foundation.
- Kurihara, K., and Sato, T. 2004. An application of the variational Bayesian Approach to Probabilistic Context-Free Grammars. In: *Proceedings of IJCNLP*. Banff, AB.

- Kurihara, K., Welling, M., and Teh, M. Y. W. 2007. Collapsed Variational Dirichlet Process Mixture Models. In: *Proceedings of IJCAI*. Hyderabad, India.
- Kuriki, S., and Takemura, A. 2001. Tail Probabilities of the Maxima of Multilinear Forms and Their Applications. *Annals of Statistics*, **29**(2), 328–371.
- Lee, T. L., Li, T. Y., and Tsai, C. H. 2008. HOM4PS-2.0: A Software Package for Solving Polynomial Systems by the Polyhedral Homotopy Continuation Method. *Computing*, **83**, 109–133.
- Levin, E., Tishby, N., and Solla, S. A. 1990. A Statistical Approaches to Learning and Generalization in Layered Neural Networks. Pages 1568–1674 of: *Proceedings of IEEE*, vol. 78.
- Li, F.-F., and Perona, P. 2005. A Bayesian Hierarchical Model for Learning Natural Scene Categories. Pages 524–531 of: *Proceedings of CVPR*. San Diego, CA.
- Lim, Y. J., and Teh, Y. W. 2007. Variational Bayesian Approach to Movie Rating Prediction. In: *Proceedings of KDD Cup and Workshop*. New York and San Jose, CA: Association for Computing Machinery.
- Lin, Z., Chen, M., and Ma, Y. 2009. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *UIUC Technical Report UILU-ENG-09-2215*.
- Liu, G., and Yan, S. 2011. Latent Low-Rank Representation for Subspace Segmentation and Feature Extraction. In: *Proceedings of ICCV*. Barcelona, Spain.
- Liu, G., Lin, Z., and Yu, Y. 2010. Robust Subspace Segmentation by Low-Rank Representation. Pages 663–670 of: *Proceedings of ICML*. Haifa, Israel: Omnipress.
- Liu, G., Xu, H., and Yan, S. 2012. Exact Subspace Segmentation and Outlier Detection by Low-Rank Representation. In: *Proceedings of AISTATS*. La Palma, Canary Islands: Proceedings of Machine Learning Research.
- Liu, X., Pasarica, C., and Shao, Y. 2003. Testing Homogeneity in Gamma Mixture Models. *Scandinavian Journal of Statistics*, **30**, 227–239.
- Lloyd, S. P. 1982. Least Square Quantization in PCM. *IEEE Transactions on Information Theory*, **28**(2), 129–137.
- MacKay, D. J. C. 1992. Bayesian Interpolation. *Neural Computation*, **4**(2), 415–447.
- MacKay, D. J. C. 1995. Developments in Probabilistic Modeling with Neural Networks—Ensemble Learning. Pages 191–198 of: *Proceedings of the 3rd Annual Symposium on Neural Networks*.
- Mackay, D. J. C. 2001. Local Minima, Symmetry-Breaking, and Model Pruning in Variational Free Energy Minimization. Available from www.inference.phy.cam.ac.uk/mackay/minima.pdf.
- MacKay, D. J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press. Available from www.inference.phy.cam.ac.uk/mackay/itila/.
- MacQueen, J. B. 1967. Some Methods for Classification and Analysis of Multivariate Observations. Pages 281–297 of: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley: University of California Press.
- Marčenko, V. A., and Pastur, L. A. 1967. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mathematics of the USSR-Sbornik*, **1**(4), 457–483.

- Marshall, A. W., Olkin, I., and Arnold, B. C. 2009. *Inequalities: Theory of Majorization and Its Applications*, 2d ed. Springer.
- Minka, T. P. 2001a. Automatic Choice of Dimensionality for PCA. Pages 598–604 of: *Advances in NIPS*, vol. 13. Cambridge, MA: MIT Press.
- Minka, T. P. 2001b. Expectation Propagation for Approximate Bayesian Inference. Pages 362–369 of: *Proceedings of UAI*. Seattle, WA: Morgan Kaufmann Publishers Inc.
- Mørup, M., and Hansen, L. R. 2009. Automatic Relevance Determination for Multi-Way Models. *Journal of Chemometrics*, **23**, 352–363.
- Nakajima, S., and Sugiyama, M. 2011. Theoretical Analysis of Bayesian Matrix Factorization. *Journal of Machine Learning Research*, **12**, 2579–2644.
- Nakajima, S., and Sugiyama, M. 2014. Analysis of Empirical MAP and Empirical Partially Bayes: Can They Be Alternatives to Variational Bayes? Pages 20–28 of: *Proceedings of International Conference on Artificial Intelligence and Statistics*, vol. 33. Reykjavik, Iceland: Proceedings of Machine Learning Research.
- Nakajima, S., and Watanabe, S. 2007. Variational Bayes Solution of Linear Neural Networks and Its Generalization Performance. *Neural Computation*, **19**(4), 1112–1153.
- Nakajima, S., Sugiyama, M., and Babacan, S. D. 2011 (June 28–July 2). On Bayesian PCA: Automatic Dimensionality Selection and Analytic Solution. Pages 497–504 of: *Proceedings of 28th International Conference on Machine Learning (ICML2011)*. Bellevue, WA: Omnipress.
- Nakajima, S., Sugiyama, M., Babacan, S. D., and Tomioka, R. 2013a. Global Analytic Solution of Fully-Observed Variational Bayesian Matrix Factorization. *Journal of Machine Learning Research*, **14**, 1–37.
- Nakajima, S., Sugiyama, M., and Babacan, S. D. 2013b. Variational Bayesian Sparse Additive Matrix Factorization. *Machine Learning*, **92**, 319–1347.
- Nakajima, S., Takeda, A., Babacan, S. D., Sugiyama, M., and Takeuchi, I. 2013c. Global Solver and Its Efficient Approximation for Variational Bayesian Low-Rank Subspace Clustering. In: *Advances in Neural Information Processing Systems 26*. Lake Tahoe, NV: NIPS Foundation.
- Nakajima, S., Sato, I., Sugiyama, M., Watanabe, K., and Kobayashi, H. 2014. Analysis of Variational Bayesian Latent Dirichlet Allocation: Weaker Sparsity Than MAP. Pages 1224–1232 of: *Advances in NIPS*, vol. 27. Montreal, Quebec: NIPS Foundation.
- Nakajima, S., Tomioka, R., Sugiyama, M., and Babacan, S. D. 2015. Condition for Perfect Dimensionality Recovery by Variational Bayesian PCA. *Journal of Machine Learning Research*, **16**, 3757–3811.
- Nakamura, F., and Watanabe, S. 2014. Asymptotic Behavior of Variational Free Energy for Normal Mixtures Using General Dirichlet Distribution (in Japanese). *IEICE Transactions on Information and Systems*, **J97-D(5)**, 1001–1013.
- Neal, R. M. 1996. *Bayesian Learning for Neural Networks*. New York: Springer.
- Opper, M., and Winther, O. 1996. A Mean Field Algorithm for Bayes Learning in Large Feed-Forward Neural Networks. Pages 225–231 of: *Advances in NIPS*. Denver, CO: NIPS Foundation.

- Pearson, K. 1914. *Tables for Statisticians and Biometricalians*. Cambridge: Cambridge University Press.
- Purushotham, S., Liu, Y., and Kuo, C. C. J. 2012. Collaborative Topic Regression with Social Matrix Factorization for Recommendation Systems. In: *Proceedings of ICML*. Edinburgh, UK: Omnipress.
- Rabiner, L. R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Pages 257–286 of: *Proceedings of the IEEE*. Piscataway, NJ: IEEE.
- Ranganath, R., Gerrish, S., and Blei, D. M. 2013. Black Box Variational Inference. In: *Proceedings of AISTATS*. Scottsdale, AZ: Proceedings of Machine Learning Research.
- Reinsel, G. R., and Velu, R. P. 1998. *Multivariate Reduced-Rank Regression: Theory and Applications*. New York: Springer.
- Rissanen, J. 1986. Stochastic Complexity and Modeling. *Annals of Statistics*, **14**(3), 1080–1100.
- Robbins, H., and Monro, S. 1951. A Stochastic Approximation Method. *Annals of Mathematical Statistics*, **22**(3), 400–407.
- Ruhe, A. 1970. Perturbation Bounds for Means of Eigenvalues and Invariant Subspaces. *BIT Numerical Mathematics*, **10**, 343–354.
- Rusakov, D., and Geiger, D. 2005. Asymptotic Model Selection for Naive Bayesian Networks. *Journal of Machine Learning Research*, **6**, 1–35.
- Sakamoto, T., Ishiguro, M., and Kitagawa, G. 1986. *Akaike Information Criterion Statistics*. Dordrecht: D. Reidel Publishing Company.
- Salakhutdinov, R., and Mnih, A. 2008. Probabilistic Matrix Factorization. Pages 1257–1264 of: Platt, J. C., Koller, D., Singer, Y., and Roweis, S. (eds), *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press.
- Sato, I., Kurihara, K., and Nakagawa, H. 2012. Practical Collapsed Variational Bayes Inference for Hierarchical Dirichlet Process. Pages 105–113. In: *Proceedings of KDD*. New York and Beijing, China: Association for Computing Machinery.
- Sato, M., Yoshioka, T., Kajihara, S., et al. 2004. Hierarchical Bayesian Estimation for MEG Inverse Problem. *NeuroImage*, **23**, 806–826.
- Schwarz, G. 1978. Estimating the Dimension of a Model. *Annals of Statistics*, **6**(2), 461–464.
- Seeger, M. 2008. Bayesian Inference and Optimal Design for the Sparse Linear Model. *Journal of Machine Learning Research*, **9**, 759–813.
- Seeger, M. 2009. Sparse Linear Models: Variational Approximate Inference and Bayesian Experimental Design. In: *Journal of Physics: Conference Series*, vol. 197. Bristol, UK: IOP Publishing.
- Seeger, M., and Bouchard, G. 2012. Fast Variational Bayesian Inference for Non-Conjugate Matrix Factorization Models. Pages 1012–1018. In: *Proceedings of International Conference on Artificial Intelligence and Statistics*. La Palma, Canary Islands: Proceedings of Machine Learning Research.
- Shi, J., and Malik, J. 2000. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 888–905.

- Soltanolkotabi, M., and Candès, E. J. 2011. A Geometric Analysis of Subspace Clustering with Outliers. *CoRR*. arXiv:1112.4258 [cs.IT].
- Spall, J. 2003. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. New York: John Wiley and Sons.
- Srebro, N., and Jaakkola, T. 2003. Weighted Low Rank Approximation. In: Fawcett, T., and Mishra, N. (eds), *Proceedings of the Twentieth International Conference on Machine Learning*. Washington, DC: AAAI Press.
- Srebro, N., Rennie, J., and Jaakkola, T. 2005. Maximum Margin Matrix Factorization. In: *Advances in Neural Information Processing Systems 17*. Vancouver, BC: NIPS Foundation.
- Stein, C. 1956. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. Pages 197–206 of: *Proceedings of the 3rd Berkeley Symposium on Mathematics Statistics and Probability*. Berkeley: University of California Press.
- Takemura, A., and Kuriki, S. 1997. Weights of Chi-Squared Distribution for Smooth or Piecewise Smooth Cone Alternatives. *Annals of Statistics*, **25**(6), 2368–2387.
- Teh, Y. W., Newman, D., and Welling, M. 2007. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In: *Advances in NIPS*. Vancouver, BC: NIPS Foundation.
- Tipping, M. E. 2001. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, **1**, 211–244.
- Tipping, M. E., and Bishop, C. M. 1999. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society*, **61**, 611–622.
- Tomioka, R., Suzuki, T., Sugiyama, M., and Kashima, H. 2010. An Efficient and General Augmented Lagrangian Algorithm for Learning Low-Rank Matrices. In: *Proceedings of International Conference on Machine Learning*. Haifa, Israel: Omnipress.
- Tron, R., and Vidal, R. 2007. A Benchmark for the Comparison of 3-D Motion Segmentation Algorithms. Pages 1–8. In: *Proceedings of CVPR*. Minneapolis, MN.
- Tucker, L. R. 1996. Some Mathematical Notes on Three-Mode Factor Analysis. *Psychometrika*, **31**, 279–311.
- Ueda, N., Nakano, R., Ghahramani, Z., and Hinton, G. E. 2000. SMEM Algorithm for Mixture Models. *Neural Computation*, **12**(9), 2109–2128.
- van der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge and New York: Cambridge University Press.
- Vidal, R., and Favaro, P. 2014. Low Rank Subspace Clustering. *Pattern Recognition Letters*, **43**(1), 47–61.
- Wachter, K. W. 1978. The Strong Limits of Random Matrix Spectra for Sample Matrices of Independent Elements. *Annals of Probability*, **6**, 1–18.
- Wainwright, M. J., and Jordan, M. I. 2008. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, **1**, 1–305.
- Watanabe, K. 2012. An Alternative View of Variational Bayes and Asymptotic Approximations of Free Energy. *Machine Learning*, **86**(2), 273–293.

- Watanabe, K., and Watanabe, S. 2004. Lower Bounds of Stochastic Complexities in Variational Bayes Learning of Gaussian Mixture Models. Pages 99–104 of: *Proceedings of IEEE on CIS*.
- Watanabe, K., and Watanabe, S. 2005. Variational Bayesian Stochastic Complexity of Mixture Models. Pages 99–104. In: *Advances in NIPS*, vol. 18. Vancouver, BC: NIPS Foundation.
- Watanabe, K., and Watanabe, S. 2006. Stochastic Complexities of Gaussian Mixtures in Variational Bayesian Approximation. *Journal of Machine Learning Research*, **7**, 625–644.
- Watanabe, K., and Watanabe, S. 2007. Stochastic Complexities of General Mixture Models in Variational Bayesian Learning. *Neural Networks*, **20**(2), 210–219.
- Watanabe, K., Shiga, M., and Watanabe, S. 2006. Upper Bounds for Variational Stochastic Complexities of Bayesian Networks. Pages 139–146 of: *Proceedings of IDEAL*. Burgos, Spain: Springer.
- Watanabe, K., Shiga, M., and Watanabe, S. 2009. Upper Bound for Variational Free Energy of Bayesian Networks. *Machine Learning*, **75**(2), 199–215.
- Watanabe, K., Okada, M., and Ikeda, K. 2011. Divergence Measures and a General Framework for Local Variational Approximation. *Neural Networks*, **24**(10), 1102–1109.
- Watanabe, S. 2001a. Algebraic Analysis for Nonidentifiable Learning Machines. *Neural Computation*, **13**(4), 899–933.
- Watanabe, S. 2001b. Algebraic Information Geometry for Learning Machines with Singularities. Pages 329–336 of: *Advances in NIPS*, vol. 13. Vancouver, BC: NIPS Foundation.
- Watanabe, S. 2009. *Algebraic Geometry and Statistical Learning Theory*. Cambridge: Cambridge University Press.
- Watanabe, S. 2010. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, **11**, 3571–3594.
- Watanabe, S. 2013. A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, **14**, 867–897.
- Watanabe, S., and Amari, S. 2003. Learning Coefficients of Layered Models When the True Distribution Mismatches the Singularities. *Neural Computation*, **15**, 1013–1033.
- Wei, X., and Croft, W. B. 2006. LDA-Based Document Models for Ad-Hoc Retrieval. Pages 178–185 of: *Proceedings of SIGIR*. Seattle, WA: Association for Computing Machinery New York.
- Wingate, D., and Weber, T. 2013. Automated Variational Inference in Probabilistic Programming. *arXiv:1301.1299*.
- Yamazaki, K. 2016. Asymptotic Accuracy of Bayes Estimation for Latent Variables with Redundancy. *Machine Learning*, **102**(1), 1–28.
- Yamazaki, K., and Kaji, D. 2013. Comparing Two Bayes Methods Based on the Free Energy Functions in Bernoulli Mixtures. *Neural Networks*, **44**, 36–43.
- Yamazaki, K., and Watanabe, S. 2003a. Singularities in Mixture Models and Upper Bounds Pages 1–8. of Stochastic Complexity. *Neural Networks*, **16**(7), 1029–1038.

- Yamazaki, K., and Watanabe, S. 2003b. Stochastic Complexity of Bayesian Networks. Pages 592–599 of: *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*. Acapulco, Mexico: Morgan Kaufmann.
- Yamazaki, K., and Watanabe, S. 2004. Newton Diagram and Stochastic Complexity in Mixture of Binomial Distributions. Pages 350–364. In: *Proceedings of ALT*. Padova, Italy: Springer.
- Yamazaki, K., and Watanabe, S. 2005. Algebraic Geometry and Stochastic Complexity of Hidden Markov Models. *Neurocomputing*, **69**, 62–84.
- Yamazaki, K., Aoyagi, M., and Watanabe, S. 2010. Asymptotic Analysis of Bayesian Generalization Error with Newton Diagram. *Neural Networks*, **23**(1), 35–43.

Subject Index

- N*-mode tensor, 80
 ℓ_1 -norm, 94, 291
 ℓ_1 -regularizer, 88
n-mode tensor product, 80
- activation function, 47
admissibility, 518
Akaike's information criterion (AIC), 365
all temperatures method, 382
approximate global variational Bayesian solver (AGVBS), 269
ARD Tucker, 332
asymptotic normality, 342, 352
asymptotic notation, 344
automatic relevance determination (ARD), 10, 36, 72, 204, 243, 294
average generalization error, 347
average training error, 347
- Bachmann–Landau notation, 344
backfitting algorithm, 283
basis selection effect, 373, 426
Bayes free energy, 36, 194, 348
Bayes generalization loss, 379
Bayes posterior, 4
Bayes theorem, 4
Bayes training loss, 379
Bayesian estimator, 5
Bayesian information criterion (BIC), 366
Bayesian learning, 3, 5
Bayesian network, 115, 455
Bernoulli distribution, 11, 27, 248, 253
Bernoulli mixture model, 451
Beta distribution, 11, 27
Beta function, 27
binomial distribution, 11, 27
- black-box variational inference, 50
burn-in, 59
- calculus of variations, 44
centering, 73
central limit theorem, 344
chi-squared distribution, 356
classification, 47
collaborative filtering (CF), 74, 335
collapsed Gibbs sampling, 53
collapsed MAP learning, 53
collapsed variational Bayesian learning, 53
complete likelihood, 8
conditional conjugacy, 39, 42
conditional distribution, 3
conditionally conjugate prior, 42
conjugacy, 10, 12
consistency, 352
continuation method, 267
convergence in distribution, 344
convergence in law, 344
convergence in probability, 344
coordinate descent, 66
core tensor, 80
Cramér–Rao lower-bound, 348
credible interval, 32
cross-validation, 9
cross-covariance, 73
- density, 522
Dirac delta function, 37, 52
direct site bounding, 49, 132
Dirichlet distribution, 11, 26
Dirichlet process prior, 112
distinct signal assumption, 419

- domination, 38, 518
doubly stochastic, 154
- efficiency, 518
empirical Bayesian (EBayes) estimator, 38, 193, 516
empirical Bayesian (EBayes) learning, 9, 35
empirical entropy, 349
empirical MAP (EMAP) learning, 311
empirical PB (EPB) learning, 311
empirical variational Bayesian (EVB) learning, 47
entropy, 380, 520
equivalence class, 187
error function, 47
Euler–Lagrange equation, 44
evidence, 36
evidence lower-bound (ELBO), 40, 341
exact global variational Bayesian solver (EGVBS), 267
expectation propagation (EP), 53
expectation–maximization (EM) algorithm, 9, 53, 105
exponential family, 15, 108, 443
- factor matrix, 80
finite mixture model, 103
Fisher information, 51, 197, 342, 520
foreground/background video separation, 97
forward–backward algorithm, 122
free energy, 40
free energy coefficient, 349
- Gamma distribution, 11, 16
Gamma function, 525
Gauss–Wishart distribution, 21
Gaussian distribution, 11
Gaussian mixture model, 104, 434
generalization coefficient, 348
generalization error, 335, 347
generalized Bayesian learning, 378
generalized posterior distribution, 378
generalized predictive distribution, 379
Gibbs generalization loss, 379
Gibbs learning, 380
Gibbs sampling, 59
Gibbs training loss, 379
global latent variable, 7, 103
- Hadamard product, 154
hard assignment, 9
hidden Markov model, 119, 461
hidden variable, 6
hierarchical model, 17
histogram, 26
homotopy method, 267
hyperparameter, 9
hyperprior, 9
- identifiability, 342, 352
improper prior, 200
independent and identically distributed (i.i.d.), 4
information criterion, 364
inside–outside algorithm, 126
integration effect, 374, 427
inverse temperature parameter, 379
isotropic Gauss–Gamma distribution, 17
isotropic Gaussian distribution, 11
iterative singular value shrinkage, 248, 252
- James–Stein (JS) estimator, 38, 516
Jeffreys prior, 198, 522
joint distribution, 3
- Kronecker delta, 130
Kronecker product, 66, 81, 331
Kronecker product covariance approximation (KPCA), 93, 274
Kullback–Leibler (KL) divergence, 39, 197, 347, 520
- Laplace approximation (LA), 51, 230
large-scale limit, 215, 319
latent Dirichlet allocation, 26, 127, 470
latent variable, 6
latent variable model, 103, 429
law of large numbers, 344
likelihood ratio, 347
linear neural network, 385
linear regression model, 22
link function, 253
local latent variable, 7, 103
local variational approximation, 49, 132
local-EMAP estimator, 317
local-EPB estimator, 317
local-EVB estimator, 182, 231, 242
log marginal likelihood, 36
log-concave distribution, 218
logistic regression, 132

- low-rank representation, 88
 low-rank subspace clustering (LRSC), 88, 255
- Marčenko–Pastur (MP) distribution, 215
 Marčenko–Pastur upper limit (MPUL), 216, 319
- marginal likelihood, 5
 Markov chain Monte Carlo (MCMC), 58
 matrix factorization (MF), 63, 195
 matrix variate Gaussian, 66
 maximum a posteriori (MAP) estimator, 188
 maximum a posteriori (MAP) learning, 5, 294
 maximum likelihood (ML) estimator, 188, 432
 maximum likelihood (ML) learning, 5, 105
 maximum log-likelihood, 366
 mean update (MU) algorithm, 283
 mean value theorem, 353
 metric, 522
 Metropolis–Hastings sampling, 58
 minimum description length (MDL), 366
 mixing weight, 7
 mixture coefficient, 7
 mixture model, 26, 196
 mixture of Gaussians, 104
 model distribution, 3
 model likelihood, 3
 model parameter, 3
 model selection, 364
 model-induced regularization (MIR), 72, 89, 94, 184, 195, 285, 308, 344, 373, 427
 moment matching, 54
 multilayer neural network, 196
 multinomial distribution, 11, 26
 multinomial parameter, 26
- natural parameter, 15, 108
 neural network, 47
 Newton–Raphson method, 108
 noise variance parameter, 22
 noninformative prior, 522
 nonsingular, 23
 normalization constant, 10
 normalized cuts, 88
- Occam’s razor, 201, 366
 one-of- K representation, 8, 104, 455
 overfitting, 420
 overlap (OL) method, 230, 242
- Parafac, 80
 partially Bayesian (PB) learning, 51, 131, 294
- partitioned-and-rearranged (PR) matrix, 95, 279
 plug-in predictive distribution, 6, 46
 Poisson distribution, 248, 253
 polygamma function, 108
 polynomial system, 162, 257, 267
 positive-part James–Stein (PJS) estimator, 185, 307, 390, 518
 posterior covariance, 5
 posterior distribution, 4
 posterior mean, 5
 predictive distribution, 6
 prior distribution, 3
 probabilistic context-free grammar, 123, 466
 probabilistic latent semantic analysis (pLSA), 131
 probabilistic principal component analysis (probabilistic PCA), 63, 71, 230
- quasiconvexity, 207, 305
- radial basis function (RBF), 367
 random matrix theory, 214, 375, 404
 realizability, 346, 352
 realizable, 375
 rectified linear unit (ReLU), 47
 reduced rank regression (RRR), 72, 385
 regression parameter, 22
 regular learning theory, 351
 regular model, 198, 342
 regularity condition, 342, 351
 relative Bayes free energy, 348
 relative variational Bayesian (VB) free energy, 383
 resolution of singularities, 378
 robust principal component analysis (robust PCA), 93, 288
- sample mean, 13
 score function, 50
 segmentation-based SAMF (sSAMF), 289
 selecting the optimal basis function, 372
 self-averaging, 216
 sigmoid function, 47, 248, 253
 simple variational Bayesian (SimpleVB) learning, 71
 singular learning theory (SLT), 376
 singular model, 197, 342, 522
 singularities, 197, 342
 soft assignment, 9

- sparse additive matrix factorization (SAMF), 94, 96, 204, 279
sparse matrix factorization (SMF) term, 94, 204, 279
sparse subspace clustering, 88
sparsity-inducing prior, 135
spectral clustering algorithm, 88
spiked covariance (SC) distribution, 217
spiked covariance model, 214
standard $(K - 1)$ -simplex, 7
state density, 376
stick-breaking process, 112
stochastic complexity, 36
stochastic gradient descent, 50
strictly quasiconnex, 207
strong unimodality, 218
subspace clustering, 87
subtle signal assumption, 420
sufficient statistics, 15
superdiagonal, 80
- Taylor approximation, 51
tensor, 80
tensor mode, 80
tensor rank, 80
trace norm, 88, 94, 291
training coefficient, 348
- training error, 347
trial distribution, 39
Tucker factorization (TF), 80, 294, 331, 336
- underfitting, 420
unidentifiability, 184, 187, 195
uniform prior, 198, 522
unnormalized posterior distribution, 4
- variation, 44
variational Bayesian (VB) estimator, 46
variational Bayesian (VB) free energy, 383
variational Bayesian (VB) learning, 39
variational Bayesian (VB) posterior, 43, 46
variational parameter, 46, 66
vectorization operator, 66, 81, 331
volume element, 197, 522
- weak convergence, 344
whitening, 73, 386
widely applicable Bayesian information criterion (WBIC), 60, 381
widely applicable information criterion (WAIC), 380
Wishart distribution, 11, 20
- zeta function, 376