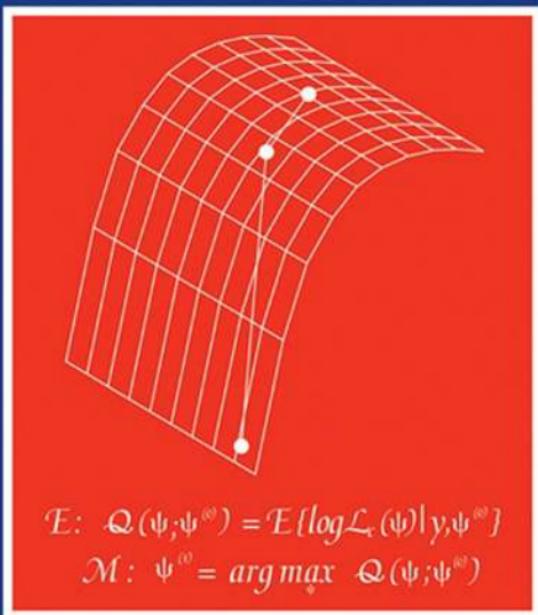


The EM Algorithm and Extensions

Second Edition



Geoffrey J. McLachlan
Thriyambakam Krishnan

www.
LINK AVAILABLE

The EM Algorithm and Extensions

Second Edition

Geoffrey J. McLachlan

The University of Queensland

*Department of Mathematics and Institute for Molecular Bioscience
St. Lucia, Australia*

Thriyambakam Krishnan

*Cranes Software International Limited
Bangalore, India*



A JOHN WILEY & SONS, INC., PUBLICATION

This Page Intentionally Left Blank

The EM Algorithm and Extensions

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice,
Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith,
Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.

The EM Algorithm and Extensions

Second Edition

Geoffrey J. McLachlan

The University of Queensland

*Department of Mathematics and Institute for Molecular Bioscience
St. Lucia, Australia*

Thriyambakam Krishnan

*Cranes Software International Limited
Bangalore, India*



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2008 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

McLachlan, Geoffrey J., 1946-

The EM algorithm and extensions / Geoffrey J. McLachlan,
Thriyambakam Krishnan. — 2nd ed.

p. cm.

ISBN 978-0-471-20170-0 (cloth)

1. Expectation-maximization algorithms. 2. Estimation theory. 3.

Missing observations (Statistics) I. Krishnan, T.

(Thriyambakam), 193— II. Title.

QA276.8.M394 2007

519.5'44—dc22

2007017908

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To
Beryl, Jonathan, and Robbie

This Page Intentionally Left Blank

CONTENTS

PREFACE TO THE SECOND EDITION	xix
PREFACE TO THE FIRST EDITION	xxi
LIST OF EXAMPLES	xxv
1 GENERAL INTRODUCTION	1
1.1 Introduction	1
1.2 Maximum Likelihood Estimation	3
1.3 Newton-Type Methods	5
1.3.1 Introduction	5
1.3.2 Newton-Raphson Method	5
1.3.3 Quasi-Newton Methods	6
1.3.4 Modified Newton Methods	6
1.4 Introductory Examples	8
1.4.1 Introduction	8
1.4.2 Example 1.1: A Multinomial Example	8

vii

1.4.3	Example 1.2: Estimation of Mixing Proportions	13
1.5	Formulation of the EM Algorithm	18
1.5.1	EM Algorithm	18
1.5.2	Example 1.3: Censored Exponentially Distributed Survival Times	20
1.5.3	E- and M-Steps for the Regular Exponential Family	22
1.5.4	Example 1.4: Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	23
1.5.5	Generalized EM Algorithm	24
1.5.6	GEM Algorithm Based on One Newton-Raphson Step	24
1.5.7	EM Gradient Algorithm	25
1.5.8	EM Mapping	26
1.6	EM Algorithm for MAP and MPL Estimation	26
1.6.1	Maximum <i>a Posteriori</i> Estimation	26
1.6.2	Example 1.5: A Multinomial Example (<i>Example 1.1 Continued</i>)	27
1.6.3	Maximum Penalized Estimation	27
1.7	Brief Summary of the Properties of the EM Algorithm	28
1.8	History of the EM Algorithm	29
1.8.1	Early EM History	29
1.8.2	Work Before Dempster, Laird, and Rubin (1977)	29
1.8.3	EM Examples and Applications Since Dempster, Laird, and Rubin (1977)	31
1.8.4	Two Interpretations of EM	32
1.8.5	Developments in EM Theory, Methodology, and Applications	
	33	
1.9	Overview of the Book	36
1.10	Notations	37

2	EXAMPLES OF THE EM ALGORITHM	41
2.1	Introduction	41
2.2	Multivariate Data with Missing Values	42
2.2.1	Example 2.1: Bivariate Normal Data with Missing Values	42
2.2.2	Numerical Illustration	45
2.2.3	Multivariate Data: Buck's Method	45
2.3	Least Squares with Missing Data	47
2.3.1	Healy–Westmacott Procedure	47

2.3.2	Example 2.2: Linear Regression with Missing Dependent Values	47
2.3.3	Example 2.3: Missing Values in a Latin Square Design	49
2.3.4	Healy–Westmacott Procedure as an EM Algorithm	49
2.4	Example 2.4: Multinomial with Complex Cell Structure	51
2.5	Example 2.5: Analysis of PET and SPECT Data	54
2.6	Example 2.6: Multivariate t -Distribution (Known D.F.)	58
2.6.1	ML Estimation of Multivariate t -Distribution	58
2.6.2	Numerical Example: Stack Loss Data	61
2.7	Finite Normal Mixtures	61
2.7.1	Example 2.7: Univariate Component Densities	61
2.7.2	Example 2.8: Multivariate Component Densities	64
2.7.3	Numerical Example: Red Blood Cell Volume Data	65
2.8	Example 2.9: Grouped and Truncated Data	66
2.8.1	Introduction	66
2.8.2	Specification of Complete Data	66
2.8.3	E-Step	69
2.8.4	M-Step	70
2.8.5	Confirmation of Incomplete-Data Score Statistic	70
2.8.6	M-Step for Grouped Normal Data	71
2.8.7	Numerical Example: Grouped Log Normal Data	72
2.9	Example 2.10: A Hidden Markov AR(1) model	73
3	BASIC THEORY OF THE EM ALGORITHM	77
3.1	Introduction	77
3.2	Monotonicity of the EM Algorithm	78
3.3	Monotonicity of a Generalized EM Algorithm	79
3.4	Convergence of an EM Sequence to a Stationary Value	79
3.4.1	Introduction	79
3.4.2	Regularity Conditions of Wu (1983)	80
3.4.3	Main Convergence Theorem for a Generalized EM Sequence	81
3.4.4	A Convergence Theorem for an EM Sequence	82
3.5	Convergence of an EM Sequence of Iterates	83
3.5.1	Introduction	83
3.5.2	Two Convergence Theorems of Wu (1983)	83
3.5.3	Convergence of an EM Sequence to a Unique Maximum Likelihood Estimate	84

3.5.4	Constrained Parameter Spaces	84
3.6	Examples of Nontypical Behavior of an EM (GEM) Sequence	85
3.6.1	Example 3.1: Convergence to a Saddle Point	85
3.6.2	Example 3.2: Convergence to a Local Minimum	88
3.6.3	Example 3.3: Nonconvergence of a Generalized EM Sequence	90
3.6.4	Example 3.4: Some E-Step Pathologies	93
3.7	Score Statistic	95
3.8	Missing Information	95
3.8.1	Missing Information Principle	95
3.8.2	Example 3.5: Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	96
3.9	Rate of Convergence of the EM Algorithm	99
3.9.1	Rate Matrix for Linear Convergence	99
3.9.2	Measuring the Linear Rate of Convergence	100
3.9.3	Rate Matrix in Terms of Information Matrices	101
3.9.4	Rate Matrix for Maximum <i>a Posteriori</i> Estimation	102
3.9.5	Derivation of Rate Matrix in Terms of Information Matrices	102
3.9.6	Example 3.6: Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	103
4	STANDARD ERRORS AND SPEEDING UP CONVERGENCE	105
4.1	Introduction	105
4.2	Observed Information Matrix	106
4.2.1	Direct Evaluation	106
4.2.2	Extraction of Observed Information Matrix in Terms of the Complete-Data Log Likelihood	106
4.2.3	Regular Case	108
4.2.4	Evaluation of the Conditional Expected Complete-Data Information Matrix	108
4.2.5	Examples	109
4.3	Approximations to Observed Information Matrix: i.i.d. Case	114
4.4	Observed Information Matrix for Grouped Data	116
4.4.1	Approximation Based on Empirical Information	116
4.4.2	Example 4.3: Grouped Data from an Exponential Distribution	117
4.5	Supplemented EM Algorithm	120

4.5.1	Definition	120
4.5.2	Calculation of $\mathbf{J}(\hat{\boldsymbol{\Psi}})$ via Numerical Differentiation	122
4.5.3	Stability	123
4.5.4	Monitoring Convergence	124
4.5.5	Difficulties of the SEM Algorithm	124
4.5.6	Example 4.4: Univariate Contaminated Normal Data	125
4.5.7	Example 4.5: Bivariate Normal Data with Missing Values	128
4.6	Bootstrap Approach to Standard Error Approximation	130
4.7	Baker's, Louis', and Oakes' Methods for Standard Error Computation	131
4.7.1	Baker's Method for Standard Error Computation	131
4.7.2	Louis' Method of Standard Error Computation	132
4.7.3	Oakes' Formula for Standard Error Computation	133
4.7.4	Example 4.6: Oakes' Standard Error for Example 1.1	134
4.7.5	Example 4.7: Louis' Method for Example 2.4	134
4.7.6	Baker's Method for Standard Error for Categorical Data	135
4.7.7	Example 4.8: Baker's Method for Example 2.4	136
4.8	Acceleration of the EM Algorithm via Aitken's Method	137
4.8.1	Aitken's Acceleration Method	137
4.8.2	Louis' Method	137
4.8.3	Example 4.9: Multinomial Data	138
4.8.4	Example 4.10: Geometric Mixture	139
4.8.5	Example 4.11: Grouped and Truncated Data. (<i>Example 2.8 Continued</i>)	142
4.9	An Aitken Acceleration-Based Stopping Criterion	142
4.10	Conjugate Gradient Acceleration of EM Algorithm	144
4.10.1	Conjugate Gradient Method	144
4.10.2	A Generalized Conjugate Gradient Algorithm	144
4.10.3	Accelerating the EM Algorithm	145
4.11	Hybrid Methods for Finding the MLE	146
4.11.1	Introduction	146
4.11.2	Combined EM and Modified Newton-Raphson Algorithm	146
4.12	A GEM Algorithm Based on One Newton-Raphson Step	148
4.12.1	Derivation of a Condition to be a Generalized EM Sequence	148
4.12.2	Simulation Experiment	149
4.13	EM Gradient Algorithm	149
4.14	A Quasi-Newton Acceleration of the EM Algorithm	151
4.14.1	The Method	151

4.14.2 Example 4.12: Dirichlet Distribution	153
4.15 Ikeda Acceleration	157
5 EXTENSIONS OF THE EM ALGORITHM	159
5.1 Introduction	159
5.2 ECM Algorithm	160
5.2.1 Motivation	160
5.2.2 Formal Definition	160
5.2.3 Convergence Properties	162
5.2.4 Speed of Convergence	162
5.2.5 Convergence Rates of EM and ECM	163
5.2.6 Example 5.1: ECM Algorithm for Hidden Markov AR(1) Model	164
5.2.7 Discussion	164
5.3 Multicycle ECM Algorithm	165
5.4 Example 5.2: Normal Mixtures with Equal Correlations	166
5.4.1 Normal Components with Equal Correlations	166
5.4.2 Application of ECM Algorithm	166
5.4.3 Fisher's <i>Iris</i> Data	168
5.5 Example 5.3: Mixture Models for Survival Data	168
5.5.1 Competing Risks in Survival Analysis	168
5.5.2 A Two-Component Mixture Regression Model	169
5.5.3 Observed Data	169
5.5.4 Application of EM Algorithm	170
5.5.5 M-Step for Gompertz Components	171
5.5.6 Application of a Multicycle ECM Algorithm	172
5.5.7 Other Examples of EM Algorithm in Survival Analysis	173
5.6 Example 5.4: Contingency Tables with Incomplete Data	174
5.7 ECME Algorithm	175
5.8 Example 5.5: MLE of t -Distribution with Unknown D.F.	176
5.8.1 Application of the EM Algorithm	176
5.8.2 M-Step	177
5.8.3 Application of ECM Algorithm	177
5.8.4 Application of ECME Algorithm	178
5.8.5 Some Standard Results	178
5.8.6 Missing Data	179
5.8.7 Numerical Examples	181

5.8.8	Theoretical Results on the Rate of Convergence	181
5.9	Example 5.6: Variance Components	182
5.9.1	A Variance Components Model	182
5.9.2	E-Step	183
5.9.3	M-Step	184
5.9.4	Application of Two Versions of ECME Algorithm	185
5.9.5	Numerical Example	185
5.10	Linear Mixed Models	186
5.10.1	Introduction	186
5.10.2	General Form of Linear Mixed Model	187
5.10.3	REML Estimation	188
5.10.4	Example 5.7: REML Estimation in a Hierarchical Random Effects Model	188
5.10.5	Some Other EM-Related Approaches to Mixed Model Estimation	191
5.10.6	Generalized Linear Mixed Models	191
5.11	Example 5.8: Factor Analysis	193
5.11.1	EM Algorithm for Factor Analysis	193
5.11.2	ECME Algorithm for Factor Analysis	196
5.11.3	Numerical Example	196
5.11.4	EM Algorithm in Principal Component Analysis	196
5.12	Efficient Data Augmentation	198
5.12.1	Motivation	198
5.12.2	Maximum Likelihood Estimation of t -Distribution	198
5.12.3	Variance Components Model	202
5.13	Alternating ECM Algorithm	202
5.14	Example 5.9: Mixtures of Factor Analyzers	204
5.14.1	Normal Component Factor Analyzers	205
5.14.2	E-step	205
5.14.3	CM-steps	206
5.14.4	t -Component Factor Analyzers	207
5.14.5	E-step	210
5.14.6	CM-steps	211
5.15	Parameter-Expanded EM (PX-EM) Algorithm	212
5.16	EMS Algorithm	213
5.17	One-Step-Late Algorithm	213
5.18	Variance Estimation for Penalized EM and OSL Algorithms	214

5.18.1	Penalized EM Algorithm	214
5.18.2	OSL Algorithm	215
5.18.3	Example 5.9: Variance of MPLE for the Multinomial (<i>Examples 1.1 and 4.1 Continued</i>)	215
5.19	Incremental EM	216
5.20	Linear Inverse Problems	217
6	MONTE CARLO VERSIONS OF THE EM ALGORITHM	219
6.1	Introduction	219
6.2	Monte Carlo Techniques	220
6.2.1	Integration and Optimization	220
6.2.2	Example 6.1: Monte Carlo Integration	221
6.3	Monte Carlo EM	221
6.3.1	Introduction	221
6.3.2	Example 6.2: Monte Carlo EM for Censored Data from Normal	223
6.3.3	Example 6.3: MCEM for a Two-Parameter Multinomial (<i>Example 2.4 Continued</i>)	224
6.3.4	MCEM in Generalized Linear Mixed Models	224
6.3.5	Estimation of Standard Error with MCEM	225
6.3.6	Example 6.4: MCEM Estimate of Standard Error for One-Parameter Multinomial (<i>Example 1.1 Continued</i>)	226
6.3.7	Stochastic EM Algorithm	227
6.4	Data Augmentation	228
6.4.1	The Algorithm	228
6.4.2	Example 6.5: Data Augmentation in the Multinomial (<i>Examples 1.1, 1.5 Continued</i>)	229
6.5	Bayesian EM	230
6.5.1	Posterior Mode by EM	230
6.5.2	Example 6.6: Bayesian EM for Normal with Semi-Conjugate Prior	231
6.6	I.I.D. Monte Carlo Algorithms	232
6.6.1	Introduction	232
6.6.2	Rejection Sampling Methods	233
6.6.3	Importance Sampling	234
6.7	Markov Chain Monte Carlo Algorithms	236
6.7.1	Introduction	236

6.7.2	Essence of MCMC	238
6.7.3	Metropolis–Hastings Algorithms	239
6.8	Gibbs Sampling	241
6.8.1	Introduction	241
6.8.2	Rao–Blackwellized Estimates with Gibbs Samples	242
6.8.3	Example 6.7: Why Does Gibbs Sampling Work?	243
6.9	Examples of MCMC Algorithms	245
6.9.1	Example 6.8: M–H Algorithm for Bayesian Probit Regression	245
6.9.2	Monte Carlo EM with MCMC	246
6.9.3	Example 6.9: Gibbs Sampling for the Mixture Problem	249
6.9.4	Example 6.10: Bayesian Probit Analysis with Data Augmentation	250
6.9.5	Example 6.11: Gibbs Sampling for Censored Normal	251
6.10	Relationship of EM to Gibbs Sampling	254
6.10.1	EM–Gibbs Sampling Connection	254
6.10.2	Example 6.12: EM–Gibbs Connection for Censored Data from Normal (<i>Example 6.11 Continued</i>)	256
6.10.3	Example 6.13: EM–Gibbs Connection for Normal Mixtures	257
6.10.4	Rate of Convergence of Gibbs Sampling and EM	257
6.11	Data Augmentation and Gibbs Sampling	258
6.11.1	Introduction	258
6.11.2	Example 6.14: Data Augmentation and Gibbs Sampling for Censored Normal (<i>Example 6.12 Continued</i>)	259
6.11.3	Example 6.15: Gibbs Sampling for a Complex Multinomial (<i>Example 2.4 Continued</i>)	260
6.11.4	Gibbs Sampling Analogs of ECM and ECME Algorithms	261
6.12	Empirical Bayes and EM	263
6.13	Multiple Imputation	264
6.14	Missing-Data Mechanism, Ignorability, and EM Algorithm	265
7	SOME GENERALIZATIONS OF THE EM ALGORITHM	269
7.1	Introduction	269
7.2	Estimating Equations and Estimating Functions	270
7.3	Quasi-Score and the Projection-Solution Algorithm	270
7.4	Expectation-Solution (ES) Algorithm	273
7.4.1	Introduction	273

7.4.2	Computational and Asymptotic Properties of the ES Algorithm	274
7.4.3	Example 7.1: Multinomial Example by ES Algorithm (<i>Example 1.1 Continued</i>)	274
7.5	Other Generalizations	275
7.6	Variational Bayesian EM Algorithm	276
7.7	MM Algorithm	278
7.7.1	Introduction	278
7.7.2	Methods for Constructing Majorizing/Minorizing Functions	279
7.7.3	Example 7.2: MM Algorithm for the Complex Multinomial (<i>Example 1.1 Continued</i>)	280
7.8	Lower Bound Maximization	281
7.9	Interval EM Algorithm	283
7.9.1	The Algorithm	283
7.9.2	Example 7.3: Interval-EM Algorithm for the Complex Multinomial (<i>Example 2.4 Continued</i>)	283
7.10	Competing Methods and Some Comparisons with EM	284
7.10.1	Introduction	284
7.10.2	Simulated Annealing	284
7.10.3	Comparison of SA and EM Algorithm for Normal Mixtures	285
7.11	The Delta Algorithm	286
7.12	Image Space Reconstruction Algorithm	287
8	FURTHER APPLICATIONS OF THE EM ALGORITHM	289
8.1	Introduction	289
8.2	Hidden Markov Models	290
8.3	AIDS Epidemiology	293
8.4	Neural Networks	295
8.4.1	Introduction	295
8.4.2	EM Framework for NNs	296
8.4.3	Training Multi-Layer Perceptron Networks	297
8.4.4	Intractability of the Exact E-Step for MLPs	300
8.4.5	An Integration of the Methodology Related to EM Training of RBF Networks	300
8.4.6	Mixture of Experts	301
8.4.7	Simulation Experiment	305
8.4.8	Normalized Mixtures of Experts	306

8.4.9	Hierarchical Mixture of Experts	307
8.4.10	Boltzmann Machine	308
8.5	Data Mining	309
8.6	Bioinformatics	310
REFERENCES		311
AUTHOR INDEX		339
SUBJECT INDEX		347

This Page Intentionally Left Blank

PREFACE TO THE SECOND EDITION

The second edition attempts to capture significant developments in EM methodology in the ten years since the publication of the first edition. The basic EM algorithm has two main drawbacks—slow convergence and lack of an in-built procedure to compute the covariance matrix of parameter estimates. Moreover, some complex problems lead to intractable E-steps, for which Monte Carlo methods have been shown to provide efficient solutions. There are many parallels and connections between the EM algorithm and Markov chain Monte Carlo algorithms, especially EM with data augmentation and Gibbs sampling. Furthermore, the key idea of the EM algorithm where a surrogate function of the log likelihood is maximized in a iterative procedure occurs in quite a few other optimization procedures as well, leading to a more general way of looking at EM as an optimization procedure.

Capturing the above developments in the second edition has led to updated, revised, and expanded versions of many sections of the first edition, and to the addition of two new chapters, one on Monte Carlo Versions of the EM Algorithm (Chapter 6) and another on Generalizations of the EM Algorithm (Chapter 7). These revisions and additions have necessitated the recasting of the first edition’s final (sixth) chapter, some sections of which have gone into the new chapters in different forms. The remaining sections with some additions form the last chapter with the modified title of “Further Applications of the EM Algorithm.”

The first edition of this book appeared twenty years after the publication of the seminal paper of Dempster, Laird, and Rubin (1977). This second edition appears just over ten

years after the first edition. Meng (2007) in an article entitled “Thirty Years of EM and Much More” points out how EM and MCMC are intimately related, and that both have been “workhorses for statistical computing”. The chapter on Monte Carlo Versions of the EM Algorithm attempts to bring out this EM–MCMC connection.

In this revised edition, we have drawn on material from Athreya, Delampady, and Krishnan (2003), Ng, Krishnan, and McLachlan (2004), and Krishnan (2004). Thanks are thus due to K.B. Athreya, M. Delampady, and Angus Ng.

We owe debts of gratitude to a number of other people for helping us prepare this edition: Ravindra Jore for the computations for the Linear Mixed Model Example; Mangalmurti Badgujar for carrying out Markov chain Monte Carlo computations with WinBugs, R, and SYSTAT; Arnab Chakraborty for reading the new chapters, pointing out errors and inadequacies, and giving valuable comments and suggestions; Ian Wood for reading Chapter 6 and providing us with valuable comments and suggestions; N.R. Chaganty for reading a draft of sections of Chapter 7 and giving useful comments; and David Hunter for reading sections of Chapters 3 and 7 and giving valuable comments and suggestions.

Lloyd Flack, Ian Wood, Vivien Challis, Sam Wang, and Richard Bean provided us with a great deal of LaTeX advice at various stages of the typesetting of the manuscript, and we owe them a great sense of gratitude. We thank too Devish Bhat for his assistance with the preparation of some of the figures.

The first author was supported by the Australian Research Council. Thanks are also due to the authors and owners of copyrighted material for permission to reproduce data, tables and figures. These are acknowledged in the appropriate pages in the text.

The web address for further information related to this book is:

<http://www.maths.uq.edu.au/~gjm/em2ed/>.

G.J. McLachlan
Brisbane

March 2008

T. Krishnan
Bangalore

PREFACE TO THE FIRST EDITION

This book deals with the Expectation–Maximization algorithm, popularly known as the EM algorithm. This is a general-purpose algorithm for maximum likelihood estimation in a wide variety of situations best described as *incomplete-data* problems. The name EM algorithm was given by Dempster, Laird, and Rubin in a celebrated paper read before the Royal Statistical Society in 1976 and published in its journal in 1977. In this paper, a general formulation of the EM algorithm was presented, its basic properties established, and many examples and applications of it provided. The idea behind the EM algorithm is intuitive and natural and so algorithms like it were formulated and applied in a variety of problems even before this paper. However, it was in this seminal paper that the ideas in the earlier papers were synthesized, a general formulation and a theory developed, and a host of traditional and non-traditional applications indicated. Since then, the EM algorithm has become a standard piece in the statistician’s repertoire. The incomplete-data situations where the EM algorithm has been successfully applied include not only evidently incomplete-data situations, where there are missing data, truncated distributions, censored or grouped observations, but also a whole variety of situations where the incompleteness of the data is not natural or evident. Thus, in some situations, it requires a certain amount of ingenuity on the part of the statistician to formulate the incompleteness in a suitable manner to facilitate the application of the EM algorithm in a computationally profitable manner. Following the paper of Dempster, Laird, and Rubin (1977), a spate of applications of the algorithm have appeared in the literature.

The EM algorithm is not without its limitations, many of which came to light in attempting to apply it in certain complex incomplete-data problems and some even in innocuously simple incomplete-data problems. However, a number of modifications and extensions of the algorithm has been developed to overcome some of these limitations. Thus there is a whole battery of EM-related algorithms and more are still being developed. The current developments are, however, in the direction of iterative simulation techniques or Markov Chain Monte Carlo methods, many of which can be looked upon as simulation-based versions of various EM-type algorithms.

Incomplete-data problems arise in all statistical contexts. Hence in these problems where maximum likelihood estimates usually have to be computed iteratively, there is the scope and need for an EM algorithm to tackle them. Further, even if there are no missing data or other forms of data incompleteness, it is often profitable to express the given problem as an incomplete-data one within an EM framework. For example, in some multiparameter problems like in random effects models, where an averaging over some parameters is to be carried out, an incomplete-data approach via the EM algorithm and its variants has been found useful. No wonder then that the EM algorithm has become an ubiquitous statistical tool, is a part of the entire spectrum of statistical methods, and has found applications in almost all fields where statistical techniques have been applied. The EM algorithm and its variants have been applied in such fields as medical imaging, dairy science, correcting census undercount, and AIDS epidemiology, to mention a few. Articles containing applications of the EM algorithm and even some with some methodological content have appeared in a variety of journals on statistical theory and methodology, statistical computing, and statistical applications in engineering, biology, medicine, social sciences, etc. Meng and Pedlow (1992) list a bibliography of over 1000 items and now there are at least 1700 publications related to the EM algorithm.

It is surprising that despite the obvious importance of the technique and its ramifications, no book on the subject has so far appeared. Indeed, many modern books dealing with some aspect of statistical estimation have at least some EM algorithm content. The books by Little and Rubin (1987), Tanner (1991, 1993), and Schafer (1996) have substantial EM algorithm content. But still, there seems to be a need for a full-fledged book on the subject. In our experience of lecturing to audiences of professional statisticians and to users of statistics, it appears that there is a definite need for a unified and complete treatment of the theory and methodology of the EM algorithm and its extensions, and their applications. The purpose of our writing this book is to fulfill this need. The various extensions of the EM algorithm due to Rubin, Meng, Liu, and others that have appeared in the last few years, have made this need even greater. Many extensions of the EM algorithm in the direction of iterative simulation have also appeared in recent years. Inclusion of these techniques in this book may have resulted in a more even-handed and comprehensive treatment of the EM algorithm and its extensions. However, we decided against it, since this methodology is still evolving and rapid developments in this area may make this material soon obsolete. So we have restricted this book to the EM algorithm and its variants and have only just touched upon the iterative simulation versions of it.

The book is aimed at theoreticians and practitioners of Statistics and its objective is to introduce to them the principles and methodology of the EM algorithm and its tremendous potential for applications. The main parts of the book describing the formulation of the EM algorithm, detailing its methodology, discussing aspects of its implementation, and illustrating its application in many simple statistical contexts, should be comprehensible to graduates with Statistics as their major subject. Throughout the book, the theory and methodology are illustrated with a number of examples. Where relevant, analytical exam-

ples are followed up with numerical examples. There are about thirty examples in the book. Some parts of the book, especially examples like factor analysis and variance components analysis, will need basic knowledge of these techniques to comprehend the full impact of the use of the EM algorithm. But our treatment of these examples is self-contained, although brief. However, these examples can be skipped without losing continuity.

Chapter 1 begins with a brief discussion of maximum likelihood (ML) estimation and standard approaches to the calculation of the maximum likelihood estimate (MLE) when it does not exist as a closed form solution of the likelihood equation. This is followed by a few examples of incomplete-data problems for which an heuristic derivation of the EM algorithm is given. The EM algorithm is then formulated and its basic terminology and notation established. The case of the regular exponential family (for the complete-data problem) for which the EM algorithm results in a particularly elegant solution, is specially treated. Throughout the treatment, the Bayesian perspective is also included by showing how the EM algorithm and its variants can be adapted to compute maximum *a posteriori* (MAP) estimates. The use of the EM algorithm and its variants in maximum penalized likelihood estimation (MPLE), a technique by which the MLE is smoothed, is also included.

Chapter 1 also gives a summary of the properties of the EM algorithm. Towards the end of Chapter 1, a comprehensive discussion of the history of the algorithm is presented, with a listing of the earlier ideas and examples upon which the general formulation is based. The chapter closes with a summary of the developments in the methodology since the Dempster et al. (1977) paper and with an indication of the range of applications of the algorithm.

In Chapter 2, a variety of examples of the EM algorithm is presented, following the general formulation in Chapter 1. These examples include missing values (in the conventional sense) in various experimental designs, the multinomial distribution with complex cell structure as used in genetics, the multivariate *t*-distribution for the provision of a robust estimate of a location parameter, Poisson regression models in a computerized image reconstruction process such as SPECT/PET, and the fitting of normal mixture models to grouped and truncated data as in the modeling of the volume of red blood cells.

In Chapter 3, the basic theory of the EM algorithm is systematically presented, and the monotonicity of the algorithm, convergence, and rates of convergence properties are established. The Generalized EM (GEM) algorithm and its properties are also presented. The principles of Missing Information and Self-Consistency are discussed. In this chapter, attention is inevitably given to mathematical details. However, mathematical details and theoretical points are explained and illustrated with the help of earlier and new examples. Readers not interested in the more esoteric aspects of the EM algorithm may only study the examples in this chapter or omit the chapter altogether without losing continuity.

In Chapter 4, two issues which have led to some criticism of the EM algorithm are addressed. The first concerns the provision of standard errors, or the full covariance matrix in multivariate situations, of the MLE obtained via the EM algorithm. One initial criticism of the EM algorithm was that it does not automatically provide an estimate of the covariance matrix of the MLE, as do some other approaches such as Newton-type methods. Hence we consider a number of methods for assessing the covariance matrix of the MLE $\hat{\Psi}$ of the parameter vector Ψ , obtained via the EM algorithm. Most of these methods are based on the observed information matrix. A coverage is given of methods such as the Supplemented EM algorithm that allow the observed information matrix to be calculated within the EM framework. The other common criticism that has been leveled at the EM algorithm is that its convergence can be quite slow. We therefore consider some methods that have been proposed for accelerating the convergence of the EM algorithm. They include methods

based on Aitken's acceleration procedure and the generalized conjugate gradient approach, and hybrid methods that switch from the EM algorithm after a few iterations to some Newton-type method. We consider also the use of the EM gradient algorithm as a basis of a quasi-Newton approach to accelerate convergence of the EM algorithm. This algorithm approximates the M-step of the EM algorithm by one Newton-Raphson step when the solution does not exist in closed form.

In Chapter 5, further modifications and extensions to the EM algorithm are discussed. The focus is on the Expectation–Conditional Maximum (ECM) algorithm and its extensions, including the Expectation–Conditional Maximum Either (ECME) and Alternating ECM (AECM) algorithms. The ECM algorithm is a natural extension of the EM algorithm in situations where the maximization process on the M-step is relatively simple when conditional on some function of the parameters under estimation. The ECM algorithm therefore replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps. These extensions of the EM algorithm typically produce an appreciable reduction in total computer time. More importantly, they preserve the appealing convergence properties of the EM algorithm, such as its monotone convergence.

In Chapter 6, very brief overviews are presented of iterative simulation techniques such as the Monte Carlo E-step, Stochastic EM algorithm, Data Augmentation, and the Gibbs sampler and their connections with the various versions of the EM algorithm. Then, a few methods such as Simulated Annealing, which are considered competitors to the EM algorithm, are described and a few examples comparing the performance of the EM algorithm with these competing methods are presented. The book is concluded with a brief account of the applications of the EM algorithm in such topical and interesting areas as Hidden Markov Models, AIDS epidemiology, and Neural Networks.

One of the authors (GJM) would like to acknowledge gratefully financial support from the Australian Research Council. Work on the book by one of us (TK) was facilitated by two visits to the University of Queensland, Brisbane, and a visit to Curtin University of Technology, Perth. One of the visits to Brisbane was under the Ethel Raybould Fellowship scheme. Thanks are due to these two Universities for their hospitality. Thanks are also due to the authors and owners of copyrighted material for permission to reproduce tables and figures. The authors also wish to thank Ramen Kar and Amiya Das of the Indian Statistical Institute, Calcutta, and Pauline Wilson of the University of Queensland, Brisbane, for their help with L^AT_EX and word processing, and Rudy Blazek of Michigan State University for assistance with the preparation of some of the figures.

G.J. McLachlan
Brisbane

November 1995

T. Krishnan
Calcutta

LIST OF EXAMPLES

Example Number	Title	Section	Page Number
1.1	A Multinomial Example	1.4.2	8
1.2	Estimation of Mixing Proportions	1.4.3	13
1.3	Censored Exponentially Distributed Survival Times	1.5.2	20
1.4	Censored Exponentially Distributed Survival Times <i>(Example 1.3 Continued)</i>	1.5.4	23
1.5	A Multinomial Example <i>(Example 1.1 Continued)</i>	1.6.2	27
2.1	Bivariate Normal Data with Missing Values	2.2.1	42
2.2	Linear Regression with Missing Dependent Values	2.3.2	47
2.3	Missing Values in a Latin Square Design	2.3.3	49
2.4	Multinomial with Complex Cell Structure	2.4	51
2.5	Analysis of PET and SPECT Data	2.5	54
2.6	Multivariate t-Distribution with Known Degrees of Freedom	2.6	58
2.7	(Finite Normal Mixtures) Univariate Component Densities	2.7.1	61
2.8	(Finite Normal Mixtures) Multivariate Component Densities	2.7.2	64
2.9	Grouped and Truncated Data	2.8	66
2.10	A hidden Markov AR(1) model	2.9	73
3.1	Convergence to a Saddle Point	3.6.1	85
3.2	Convergence to a Local Minimum	3.6.2	88
3.3	Nonconvergence of a Generalized EM Sequence	3.6.3	90
3.4	Some E-Step Pathologies	3.6.4	93

continued

Example Number	Title	Section	Page Number
3.5	Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	3.8.2	96
3.6	Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	3.9.6	103
4.1	Information Matrix for the Multinomial Example (<i>Example 1.1 Continued</i>)	4.2.5	109
4.2	(Information Matrix) Mixture of Two Univariate Normals with Known Common Variance	4.2.5	111
4.3	Grouped Data from an Exponential Distribution	4.4.2	117
4.4	(SEM) Univariate Contaminated Normal Data	4.5.6	125
4.5	(SEM) Bivariate Normal Data with Missing Values	4.5.7	128
4.6	Oakes' Standard Error for Example 1.1	4.7.4	134
4.7	Louis' Method for Example 2.4	4.7.5	134
4.8	Baker's Method for Example 2.4	4.7.7	136
4.9	(Aitken Acceleration) Multinomial Data	4.8.3	138
4.10	(Aitken Acceleration) Geometric Mixture	4.8.4	139
4.11	(Aitken Acceleration) Grouped and Truncated Data (<i>Example 2.8 Continued</i>)	4.8.5	142
4.12	(Quasi-Newton Acceleration) Dirichlet Distribution	4.14.2	153
5.1	ECM Algorithm for Hidden Markov AR(1) Model	5.2.6	164
5.2	(ECM) Normal Mixtures with Equal Correlations	5.4	166
5.3	(ECM) Mixture Models for Survival Data	5.5	168
5.4	(ECM) Contingency Tables with Incomplete Data	5.6	174
5.5	MLE of <i>t</i> -Distribution with Unknown D.F.	5.8	176
5.6	(ECME Algorithm) Variance Components	5.9	182
5.7	REML Estimation in a Hierarchical Random Effects Model	5.10.4	188
5.8	(ECM Algorithm) Factor Analysis	5.11	192
5.9	Mixtures of Factor Analyzers	5.14	204
5.10	Variance of MPLE for the Multinomial (<i>Examples 1.1 and 4.1 Continued</i>)	5.18.3	215

continued

Example Number	Title	Section	Page Number
6.1	Monte Carlo Integration	6.2.2	221
6.2	Monte Carlo EM for Censored Data from Normal	6.3.2	223
6.3	MCEM for a Two-Parameter Multinomial (<i>Example 2.4 Continued</i>)	6.3.3	224
6.4	MCEM Estimate of Standard Error for One-Parameter Multinomial (<i>Example 1.1 Continued</i>)	6.3.6	226
6.5	Data Augmentation in the Multinomial (<i>Examples 1.1, 1.5 Continued</i>)	6.4.2	229
6.6	Bayesian EM for Normal with Semi-Conjugate Prior	6.5.2	231
6.7	Why Does Gibbs Sampling Work?	6.8.3	243
6.8	M-H Algorithm for Bayesian Probit Regression	6.9.1	245
6.9	Gibbs Sampling for the Mixture Problem	6.9.3	249
6.10	Bayesian Probit Analysis with Data Augmentation	6.9.4	250
6.11	Gibbs Sampling for Censored Normal	6.9.5	251
6.12	EM–Gibbs Connection for Censored Data from Normal (<i>Example 6.11 Continued</i>)	6.10.2	256
6.13	EM-Gibbs Connection for Normal Mixtures	6.10.3	257
6.14	Data Augmentation and Gibbs Sampling for Censored Normal (<i>Example 6.12 Continued</i>)	6.11.2	259
6.15	Gibbs Sampling for a Complex Multinomial (<i>Example 2.4 Continued</i>)	6.11.3	260
7.1	Multinomial Example by ES Algorithm (<i>Example 1.1 Continued</i>)	7.4.3	274
7.2	MM Algorithm for the Complex Multinomial (<i>Example 1.1 Continued</i>)	7.7.3	280
7.3	Interval-EM Algorithm for the Complex Multinomial (<i>Example 2.4 Continued</i>)	7.9.2	283

This Page Intentionally Left Blank

CHAPTER 1

GENERAL INTRODUCTION

1.1 INTRODUCTION

The Expectation-Maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood (ML) estimates, useful in a variety of incomplete-data problems, where algorithms such as the Newton-Raphson method may turn out to be more complicated. On each iteration of the EM algorithm, there are two steps—called the *Expectation step* or the E-step and the *Maximization step* or the M-step. Because of this, the algorithm is called the EM algorithm. This name was given by Dempster, Laird, and Rubin (1977) in their fundamental paper. (We hereafter refer to this paper as the DLR paper or simply DLR.) The situations where the EM algorithm is profitably applied can be described as *incomplete-data problems*, where ML estimation is made difficult by the absence of some part of data in a more familiar and simpler data structure. The EM algorithm is closely related to the *ad hoc* approach to estimation with missing data, where the parameters are estimated after filling in initial values for the missing data. The latter are then updated by their predicted values using these initial parameter estimates. The parameters are then reestimated, and so on, proceeding iteratively until convergence.

This idea behind the EM algorithm being intuitive and natural, algorithms like the EM have been formulated and applied in a variety of problems even before the DLR paper. But it was in the seminal DLR paper that the ideas were synthesized, a general formulation of the EM algorithm established, its properties investigated, and a host of traditional and non-traditional applications indicated.

The situations where the EM algorithm can be applied include not only evidently incomplete-data situations, where there are missing data, truncated distributions, or censored or grouped observations, but also a whole variety of situations where the incompleteness of the data is not all that natural or evident. These include statistical models such as random effects, mixtures, convolutions, log linear models, and latent class and latent variable structures. Hitherto intractable ML estimation problems for these situations have been solved or complicated ML estimation procedures have been simplified using the EM algorithm. The EM algorithm has thus found applications in almost all statistical contexts and in almost all fields where statistical techniques have been applied—medical imaging, dairy science, correcting census undercount, and AIDS epidemiology, to mention a few.

Data sets with missing values, censored and grouped observations, and models with truncated distributions, etc., which result in complicated likelihood functions cannot be avoided in practical situations. The development of the EM algorithm and related methodology together with the availability of inexpensive and rapid computing power have made analysis of such data sets much more tractable than they were earlier. The EM algorithm has already become a standard tool in the statistical repertoire.

The basic idea of the EM algorithm is to associate with the given *incomplete-data problem*, a *complete-data problem* for which ML estimation is computationally more tractable; for instance, the complete-data problem chosen may yield a closed form solution to the maximum likelihood estimate (MLE) or may be amenable to MLE computation with a standard computer package. The methodology of the EM algorithm then consists in reformulating the problem in terms of this more easily solved complete-data problem, establishing a relationship between the likelihoods of these two problems, and exploiting the simpler MLE computation of the complete-data problem in the M-step of the iterative computing algorithm.

Although a problem at first sight may not appear to be an incomplete-data one, there may be much to be gained computation-wise by artificially formulating it as such to facilitate ML estimation. This is because the EM algorithm exploits the reduced complexity of ML estimation given the complete data. For many statistical problems, the complete-data likelihood has a nice form. The E-step consists in manufacturing data for the complete-data problem, using the observed data set of the incomplete-data problem and the current value of the parameters, so that the simpler M-step computation can be applied to this ‘completed’ data set. More precisely, it is the log likelihood of the complete-data problem that is “manufactured” in the E-step. As it is based partly on unobservable data, it is replaced by its conditional expectation given the observed data, where this E-step is effected using the current fit for the unknown parameters. Starting from suitable initial parameter values, the E- and M-steps are repeated until convergence. Of course, the complete-data problem is to be suitably chosen from the point of view of simplicity of the complete-data MLE’s; it may even be a hypothetical problem from the point of view of practical implementation. For instance, a complete-data problem defined in the context of factor analysis has data on unobservable latent variables.

Although the EM algorithm has been successfully applied in a variety of contexts, it can in certain situations be painfully slow to converge. This has resulted in the development of modified versions of the algorithm as well as many simulation-based methods and other extensions of it. This area is still developing. An initial criticism of the EM algorithm was that it did not produce estimates of the covariance matrix of the MLE’s. However, developments subsequent to the DLR paper have provided methods for such estimation, which can be integrated into the EM computational scheme.

In the next section, we shall provide a very brief description of ML estimation. However, in this book we do not discuss the question of why use MLE's; excellent treatises are available wherein the attractive properties of MLE's are established (for instance, Rao, 1973; Cox and Hinkley, 1974; Lehmann, 1983; Lehmann and Casella, 2003; Stuart and Ord, 1994). In Section 1.3, we briefly define the standard techniques before the advent of the EM algorithm for the computation MLE's, which are the Newton-Raphson method and its variants such as Fisher's scoring method and quasi-Newton methods. For a more detailed account of these and other numerical methods for statistical computation, the reader is referred to books that discuss these methods; for example, Everitt (1987), Thisted (1988), and Lange (1999) discuss ML estimation and Dennis and Schanbel (1983), Ratschek and Rokne (1988), and Lange (2004) discuss optimization methods in general.

1.2 MAXIMUM LIKELIHOOD ESTIMATION

Although we shall focus on the application of the EM algorithm for computing MLE's in a frequentist framework, it can be equally applied to find the mode of the posterior distribution in a Bayesian framework.

We let \mathbf{Y} be a p -dimensional random vector with probability density function (p.d.f.) $g(\mathbf{y}; \boldsymbol{\Psi})$ on \mathbb{R}^p , where $\boldsymbol{\Psi} = (\Psi_1, \dots, \Psi_d)^T$ is the vector containing the unknown parameters in the postulated form for the p.d.f. of \mathbf{Y} . Here (and also elsewhere in this book) the superscript T denotes the transpose of a vector or a matrix. The parameter space is denoted by Ω . Although we are taking \mathbf{Y} to be a continuous random vector, we can still view $g(\mathbf{y}; \boldsymbol{\Psi})$ as a p.d.f. in the case where \mathbf{Y} is discrete by the adoption of counting measure.

For example, if $\mathbf{w}_1, \dots, \mathbf{w}_n$ denotes an observed random sample of size n on some random vector \mathbf{W} with p.d.f. $f(\mathbf{w}; \boldsymbol{\Psi})$, then

$$\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T$$

and

$$g(\mathbf{y}; \boldsymbol{\Psi}) = \prod_{j=1}^n f(\mathbf{w}_j; \boldsymbol{\Psi}).$$

The vector $\boldsymbol{\Psi}$ is to be estimated by maximum likelihood. The likelihood function for $\boldsymbol{\Psi}$ formed from the observed data \mathbf{y} is given by

$$L(\boldsymbol{\Psi}) = g(\mathbf{y}; \boldsymbol{\Psi}).$$

An estimate $\hat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}$ can be obtained as a solution of the likelihood equation

$$\partial L(\boldsymbol{\Psi}/\partial \boldsymbol{\Psi}) = \mathbf{0},$$

or equivalently,

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}. \quad (1.1)$$

Briefly, the aim of ML estimation (Lehmann, 1983; Lehmann and Casella, 2003) is to determine an estimate for each n ($\hat{\boldsymbol{\Psi}}$ in the present context), so that it defines a sequence of roots of the likelihood equation that is consistent and asymptotically efficient. Such a sequence is known to exist under suitable regularity conditions (Cramér, 1946). With probability tending to one, these roots correspond to local maxima in the interior of the parameter space. For estimation models in general, the likelihood usually has a global

maximum in the interior of the parameter space. Then typically a sequence of roots of the likelihood equation with the desired asymptotic properties is provided by taking $\hat{\Psi}$ for each n to be the root that globally maximizes the likelihood; that is, $\hat{\Psi}$ is the MLE. We shall henceforth refer to $\hat{\Psi}$ as the MLE, even though it may not globally maximize the likelihood. Indeed, in some of the examples on mixture models to be presented, the likelihood is unbounded. However, for these models there may still exist under the usual regularity conditions a sequence of roots of the likelihood equation with the properties of consistency, efficiency, and asymptotic normality; see McLachlan and Basford (1988, Chapter 1). We let

$$\mathbf{I}(\Psi; \mathbf{y}) = -\partial^2 \log L(\Psi) / \partial \Psi \partial \Psi^T$$

be the matrix of the negative of the second-order partial derivatives of the log likelihood function with respect to the elements of Ψ . Under regularity conditions, the expected (Fisher) information matrix $\mathcal{I}(\Psi)$ is given by

$$\begin{aligned}\mathcal{I}(\Psi) &= E_{\Psi}\{S(\mathbf{Y}; \Psi)S^T(\mathbf{Y}; \Psi)\} \\ &= -E_{\Psi}\{\mathbf{I}(\Psi; \mathbf{Y})\},\end{aligned}$$

where

$$S(\mathbf{y}; \Psi) = \partial \log L(\Psi) / \partial \Psi$$

is the gradient vector of the log likelihood function; that is, the score statistic. Here and elsewhere in this book, the operator E_{Ψ} denotes expectation using the parameter vector Ψ .

The asymptotic covariance matrix of the MLE $\hat{\Psi}$ is equal to the inverse of the expected information matrix $\mathcal{I}(\Psi)$, which can be approximated by $\mathcal{I}(\hat{\Psi})$; that is, the standard error of $\hat{\Psi}_i = (\hat{\Psi})_i$ is given by

$$SE(\hat{\Psi}_i) \approx (\mathcal{I}^{-1}(\hat{\Psi}))_{ii}^{1/2} \quad (i = 1, \dots, d), \quad (1.2)$$

where the standard notation $(\mathbf{A})_{ij}$ is used for the element in the i th row and j th column of a matrix \mathbf{A} .

It is common in practice to estimate the inverse of the covariance matrix of a maximum solution by the observed information matrix $\mathbf{I}(\hat{\Psi}; \mathbf{y})$, rather than the expected information matrix $\mathcal{I}(\Psi)$ evaluated at $\Psi = \hat{\Psi}$. This approach gives the approximation

$$SE(\hat{\Psi}_i) \approx (\mathbf{I}^{-1}(\hat{\Psi}; \mathbf{y}))_{ii}^{1/2} \quad (i = 1, \dots, d). \quad (1.3)$$

Efron and Hinkley (1978) have provided a frequentist justification of (1.3) over (1.2) in the case of one-parameter ($d = 1$) families. Also, the observed information matrix is usually more convenient to use than the expected information matrix, as it does not require an expectation to be taken.

Often in practice the log likelihood function cannot be maximized analytically. In such cases, it may be possible to compute iteratively the MLE of Ψ by using a Newton-Raphson maximization procedure or some variant, provided the total number d of parameters in the model is not too large. Another alternative is to apply the EM algorithm. Before we proceed with the presentation of the EM algorithm, we briefly define in the next section the Newton-Raphson method and some variants for the computation of the MLE.

1.3 NEWTON-TYPE METHODS

1.3.1 Introduction

Since the properties of the EM algorithm are to be contrasted with those of Newton-type methods, which are the main alternatives for the computation of MLE's, we now give a brief review of the Newton-Raphson method and some variants.

Like many other methods for computing MLE's, the EM algorithm is a method for finding zeros of a function. In numerical analysis there are various techniques for finding zeros of a specified function, including the Newton-Raphson (NR) method, quasi-Newton methods, and modified Newton methods. In a statistical framework, the modified Newton methods include the scoring algorithm of Fisher and its modified version using the empirical information matrix in place of the expected information matrix.

1.3.2 Newton-Raphson Method

The Newton-Raphson method for solving the likelihood equation

$$\mathbf{S}(\mathbf{y}; \boldsymbol{\Psi}) = \mathbf{0}, \quad (1.4)$$

approximates the gradient vector $\mathbf{S}(\mathbf{y}; \boldsymbol{\Psi})$ of the log likelihood function $\log L(\boldsymbol{\Psi})$ by a linear Taylor series expansion about the current fit $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. This gives

$$\mathbf{S}(\mathbf{y}; \boldsymbol{\Psi}) \approx \mathbf{S}(\mathbf{y}; \boldsymbol{\Psi}^{(k)}) - \mathbf{I}(\boldsymbol{\Psi}^{(k)}; \mathbf{y})(\boldsymbol{\Psi} - \boldsymbol{\Psi}^{(k)}). \quad (1.5)$$

A new fit $\boldsymbol{\Psi}^{(k+1)}$ is obtained by taking it to be a zero of the right-hand side of (1.5). Hence

$$\boldsymbol{\Psi}^{(k+1)} = \boldsymbol{\Psi}^{(k)} + \mathbf{I}^{-1}(\boldsymbol{\Psi}^{(k)}; \mathbf{y})\mathbf{S}(\mathbf{y}; \boldsymbol{\Psi}^{(k)}). \quad (1.6)$$

If the log likelihood function is concave and unimodal, then the sequence of iterates $\{\boldsymbol{\Psi}^{(k)}\}$ converges to the MLE of $\boldsymbol{\Psi}$, in one step if the log likelihood function is a quadratic function of $\boldsymbol{\Psi}$. When the log likelihood function is not concave, the Newton-Raphson method is not guaranteed to converge from an arbitrary starting value. Under reasonable assumptions on $L(\boldsymbol{\Psi})$ and a sufficiently accurate starting value, the sequence of iterates $\boldsymbol{\Psi}^{(k)}$ produced by the Newton-Raphson method enjoys local quadratic convergence to a solution $\boldsymbol{\Psi}^*$ of (1.4). That is, given a norm $\|\cdot\|$ on Ω , there is a constant h such that if $\boldsymbol{\Psi}^{(0)}$ is sufficiently close to $\boldsymbol{\Psi}^*$, then

$$\|\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^*\| \leq h\|\boldsymbol{\Psi}^{(k)} - \boldsymbol{\Psi}^*\|^2$$

holds for $k = 0, 1, 2, \dots$. Quadratic convergence is ultimately very fast, and it is regarded as the major strength of the Newton-Raphson method. But there can be potentially severe problems with this method in applications. Firstly, it requires at each iteration, the computation of the $d \times d$ information matrix $\mathbf{I}(\boldsymbol{\Psi}^{(k)}; \mathbf{y})$ (that is, the negative of the Hessian matrix) and the solution of a system of d linear equations. In general, this is achieved at a cost of $O(d^3)$ arithmetic operations. Thus the computation required for an iteration of the Newton-Raphson method is likely to become expensive very rapidly as d becomes large. One must allow for the storage of the Hessian or some set of factors of it. Furthermore, the Newton-Raphson method in its basic form (1.6) requires for some problems an impractically accurate initial value for $\boldsymbol{\Psi}$ for the sequence of iterates $\{\boldsymbol{\Psi}^{(k)}\}$ to converge to a solution of (1.4). It has the tendency to head toward saddle points and local minima as often as toward

local maxima. In some problems, however, Böhning and Lindsay (1988) show how the Newton-Raphson method can be modified to be monotonic.

Since the Newton-Raphson method requires the evaluation of $\mathbf{I}(\Psi^{(k)}; \mathbf{y})$ on each iteration k , it immediately provides an estimate of the covariance matrix of its limiting value Ψ^* (assuming it is the MLE), through the inverse of the observed information matrix $\mathbf{I}^{-1}(\Psi^{(k)}; \mathbf{y})$. Also, if the starting value is a \sqrt{n} -consistent estimator of Ψ , then the one-step iterate $\Psi^{(1)}$ is an asymptotically efficient estimator of Ψ .

1.3.3 Quasi-Newton Methods

A broad class of methods are so-called quasi-Newton methods, for which the solution of (1.4) takes the form

$$\Psi^{(k+1)} = \Psi^k - \mathbf{A}^{-1} \mathbf{S}(\mathbf{y}; \Psi^{(k)}), \quad (1.7)$$

where \mathbf{A} is used as an approximation to the Hessian matrix. This approximation can be maintained by doing a *secant update* of \mathbf{A} at each iteration. These updates are typically effected by rank-one or rank-two changes in \mathbf{A} . Methods of this class have the advantage over Newton-Raphson method of not requiring the evaluation of the Hessian matrix at each iteration and of being implementable in ways that require only $O(d^2)$ arithmetic operations to solve the system of d linear equations corresponding to (1.6) with $\mathbf{I}(\Psi^{(k)}; \mathbf{y})$ replaced by $-\mathbf{A}$. However, the full quadratic convergence of the Newton-Raphson method is lost, as a sequence of iterates $\Psi^{(k)}$ can be shown to exhibit only local superlinear convergence to a solution Ψ^* of (1.4). More precisely, suppose that the initial value $\Psi^{(0)}$ of Ψ is sufficiently near to the solution Ψ^* and that the initial value $\mathbf{A}^{(0)}$ of \mathbf{A} is sufficiently near to the Hessian matrix at the solution, that is, $-\mathbf{I}(\Psi^*; \mathbf{y})$. Then under reasonable assumptions on the likelihood function $L(\Psi)$, it can be shown that there exists a sequence h_k that converges to zero and is such that

$$\|\Psi^{(k+1)} - \Psi^*\| \leq h_k \|\Psi^{(k)} - \Psi^*\|$$

for $k = 0, 1, 2, \dots$. For further details, the reader is referred to the excellent accounts on Newton-type methods in Redner and Walker (1984) and Lange (1999).

It can be seen that quasi-Newton methods avoid the explicit evaluation of the Hessian of the log likelihood function at every iteration, as with the Newton-Raphson method. Also, they circumvent the tendency of the Newton-Raphson method to lead to saddle points and local minima as often as local maxima by forcing the approximate Hessian to be negative definite. However, as pointed out by Lange (1995b), even with these safeguards, they still have some drawbacks in many statistical applications. In particular, they usually approximate the Hessian initially by the identity, which may be a poorly scaled approximation to the problem at hand. Hence the algorithm can wildly overshoot or undershoot the maximum of the log likelihood along the direction of the current step. This has led to some alternative methods, which we now briefly mention.

1.3.4 Modified Newton Methods

Fisher's method of scoring is a member of the class of modified Newton methods, where the observed information matrix $\mathbf{I}(\Psi^{(k)}; \mathbf{y})$ for the current fit for Ψ , is replaced by $\mathcal{I}(\Psi^{(k)})$, the expected information matrix evaluated at the current fit $\Psi^{(k)}$ for Ψ .

In practice, it is often too tedious or difficult to evaluate analytically the expectation of $\mathbf{I}(\Psi; \mathbf{Y})$ to give the expected information matrix $\mathcal{I}(\Psi)$. Indeed, in some instances,

one may not wish even to perform the prerequisite task of calculating the second-order partial derivatives of the log likelihood. In that case for independent and identically distributed (i.i.d.) data, the method of scoring can be employed with the empirical information matrix $\mathbf{I}_e(\Psi^{(k)}; \mathbf{y})$ evaluated at the current fit for Ψ . The empirical information matrix $\mathbf{I}_e(\Psi^{(k)}; \mathbf{y})$ is given by

$$\begin{aligned}\mathbf{I}_e(\Psi; \mathbf{y}) &= \sum_{j=1}^n s(\mathbf{w}_j; \Psi) s^T(\mathbf{w}_j; \Psi) \\ &\quad - n^{-1} \mathbf{S}(\mathbf{y}; \Psi) \mathbf{S}^T(\mathbf{y}; \Psi)\end{aligned}\tag{1.8}$$

where $s(\mathbf{w}_j; \Psi)$ is the score function based on the single observation \mathbf{w}_j and

$$\begin{aligned}\mathbf{S}(\mathbf{y}; \Psi) &= \partial \log L(\Psi) / \partial \Psi \\ &= \sum_{j=1}^n s(\mathbf{w}_j; \Psi)\end{aligned}$$

is the score statistic for the full sample

$$\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T.$$

On evaluation at $\Psi = \hat{\Psi}$, $\mathbf{I}_e(\hat{\Psi}; \mathbf{y})$ reduces to

$$\mathbf{I}_e(\hat{\Psi}; \mathbf{y}) = \sum_{j=1}^n s(\mathbf{w}_j; \hat{\Psi}) s^T(\mathbf{w}_j; \hat{\Psi}),\tag{1.9}$$

since $\mathbf{S}(\mathbf{y}; \hat{\Psi}) = \mathbf{0}$.

Actually, Meilijson (1989) recommends forming $\mathbf{I}_e(\Psi^{(k)}; \mathbf{y})$ by evaluating (1.8) at $\Psi = \Psi^{(k)}$, since $\mathbf{S}(\mathbf{y}; \Psi^{(k)})$ is not zero. The justification of the empirical information matrix as an approximation to either the expected information matrix $\mathcal{I}(\Psi)$ or the observed information matrix $\mathbf{I}(\hat{\Psi}; \mathbf{y})$ is to be discussed in Section 4.3.

The modified Newton method, which uses on the k th iteration the empirical information matrix $\mathbf{I}_e(\Psi^{(k)}; \mathbf{y})$ in place of $\mathcal{I}(\Psi^{(k)})$, or equivalently, the Newton-Raphson method with $\mathbf{I}(\Psi^{(k)}; \mathbf{y})$ replaced by $\mathbf{I}_e(\Psi^{(k)}; \mathbf{y})$, requires $O(nd^2)$ arithmetic operations to calculate $\mathbf{I}_e(\Psi^{(k)}; \mathbf{y})$ and $O(d^3)$ arithmetic operations to solve the system of d equations implicit in (1.4). As pointed out by Redner and Walker (1984), since the $O(nd^2)$ arithmetic operations needed to compute the empirical information matrix $\mathbf{I}_e(\Psi^{(k)}; \mathbf{y})$ are likely to be considerably less expensive than the evaluation of $\mathbf{I}(\Psi^{(k)}; \mathbf{y})$ (that is, the full Hessian matrix), the cost of computation per iteration of this method should lie between that of a quasi-Newton method employing a low-rank secant update and that of the Newton-Raphson method. Under reasonable assumptions on $L(\Psi)$, one can show that, with probability one, if a solution Ψ^* of (1.4) is sufficiently close to Ψ and if n is sufficiently large, then a sequence of iterates $\{\Psi^{(k)}\}$ generated by the method of scoring or its modified version using the empirical information matrix exhibits local linear convergence to Ψ^* . That is, there is a norm $\|\cdot\|$ on Ω and a constant h , such that

$$\|\Psi^{(k+1)} - \Psi^*\| \leq h \|\Psi^{(k)} - \Psi^*\|$$

for $k = 0, 1, 2, \dots$ whenever $\Psi^{(k)}$ is sufficiently close to Ψ^* .

It is clear that the modified Newton method using the empirical information matrix (1.9) in (1.5) is an analog of the Gauss-Newton method for nonlinear least-squares estimation. With nonlinear least-squares estimation on the basis of some observed univariate random variables, w_1, \dots, w_n , one minimizes

$$\frac{1}{2} \sum_{j=1}^n \{w_j - \mu_j(\Psi)\}^2 \quad (1.10)$$

with respect to Ψ , where

$$\mu_j(\Psi) = E_\Psi(W_j).$$

The Gauss-Newton method approximates the Hessian of (1.10) by

$$\sum_{j=1}^n \{\partial \mu_j(\Psi)/\partial \Psi\} \{\partial \mu_j(\Psi)/\partial \Psi\}^T.$$

1.4 INTRODUCTORY EXAMPLES

1.4.1 Introduction

Before proceeding to formulate the EM algorithm in its generality, we give here two simple illustrative examples.

1.4.2 Example 1.1: A Multinomial Example

We consider first the multinomial example that DLR used to introduce the EM algorithm and that has been subsequently used many times in the literature to illustrate various modifications and extensions of this algorithm. It relates to a classic example of ML estimation due to Fisher (1925) and arises in genetic models for gene frequency estimation. Hartley (1958) also gave three multinomial examples of a similar nature in proposing the EM algorithm in special circumstances.

The data relate to a problem of estimation of linkage in genetics discussed by Rao (1973, pp. 368–369). The observed data vector of frequencies

$$\mathbf{y} = (y_1, y_2, y_3, y_4)^T$$

is postulated to arise from a multinomial distribution with four cells with cell probabilities

$$\frac{1}{2} + \frac{1}{4}\Psi, \frac{1}{4}(1 - \Psi), \frac{1}{4}(1 - \Psi), \text{ and } \frac{1}{4}\Psi \quad (1.11)$$

with $0 \leq \Psi \leq 1$. The parameter Ψ is to be estimated on the basis of \mathbf{y} . In the multinomial example of DLR, the observed frequencies are

$$\mathbf{y} = (125, 18, 20, 34)^T \quad (1.12)$$

from a sample of size $n = 197$. For the Newton-Raphson and scoring methods (but not the EM algorithm), Thisted (1988, Section 4.2.6) subsequently considered the same example, but with

$$\mathbf{y} = (1997, 906, 904, 32)^T \quad (1.13)$$

from a sample of size $n = 3839$. We shall consider here the results obtained for both data sets.

The probability function $g(\mathbf{y}; \Psi)$ for the observed data \mathbf{y} is given by

$$g(\mathbf{y}; \Psi) = \frac{n!}{y_1!y_2!y_3!y_4!} \left(\frac{1}{2} + \frac{1}{4}\Psi\right)^{y_1} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{y_2} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{y_3} \left(\frac{1}{4}\Psi\right)^{y_4}.$$

The log likelihood function for Ψ is, therefore, apart from an additive term not involving Ψ ,

$$\log L(\Psi) = y_1 \log(2 + \Psi) + (y_2 + y_3) \log(1 - \Psi) + y_4 \log \Psi. \quad (1.14)$$

On differentiation of (1.14) with respect to Ψ , we have that

$$\partial \log L(\Psi) / \partial \Psi = \frac{y_1}{2 + \Psi} - \frac{y_2 + y_3}{1 - \Psi} + \frac{y_4}{\Psi} \quad (1.15)$$

and

$$\begin{aligned} I(\Psi; \mathbf{y}) &= -\partial^2 \log L(\Psi) / \partial \Psi^2 \\ &= \frac{y_1}{(2 + \Psi)^2} + \frac{y_2 + y_3}{(1 - \Psi)^2} + \frac{y_4}{\Psi^2}. \end{aligned} \quad (1.16)$$

The right-hand side of (1.15) can be rewritten as a rational function, the numerator of which is a quadratic in Ψ . One of the roots is negative, and so it is the other root that we seek.

Although the likelihood equation can be solved explicitly to find the MLE $\hat{\Psi}$ of Ψ , we shall use this example to illustrate the computation of the MLE via Newton-type methods and the EM algorithm. In a later section we shall give an example of a multinomial depending on two unknown parameters where the likelihood equation cannot be solved explicitly.

Considering the Newton-type methods of computation for the data set in Thisted (1988), it can be seen from the plot of $\log L(\Psi)$ as given by (1.14) in Figure 1.1 that a choice of starting value too close to zero or much greater than 0.1 will cause difficulty with these methods. Indeed, if the Newton-Raphson procedure is started from 0.5 (the midpoint of the admissible interval for Ψ), then it converges to the negative root of -0.4668 ; the method of scoring, however, does converge to the MLE given by $\hat{\Psi} = 0.0357$ (see Thisted, 1988, page 176).

For this problem, it is not difficult to obtain a reasonable starting value since it can be seen that an unbiased estimator of Ψ is given by

$$\begin{aligned} \tilde{\Psi} &= (y_1 - y_2 - y_3 + y_4)/n \\ &= 0.0570. \end{aligned}$$

Starting the Newton-Raphson procedure from $\Psi^{(0)} = \tilde{\Psi} = 0.0570$, leads to convergence to the MLE $\hat{\Psi}$, as shown in Table 1.1. For comparative purposes, we have also displayed the iterates for the method of scoring. The latter converges more rapidly at the start, but Newton-Raphson's quadratic convergence takes over in the last few iterations. The method of scoring uses $I(\Psi^{(k)})$ instead of $I(\Psi^{(k)}; \mathbf{y})$ on each iteration k . For this problem, the expected information $I(\Psi)$ about Ψ is given on taking the expectation of (1.16) by

$$\begin{aligned} I(\Psi) &= E_{\Psi}\{I(\Psi; \mathbf{Y})\} \\ &= \frac{n}{4} \left\{ \frac{1}{2 + \Psi} + \frac{2}{(1 - \Psi)} + \frac{1}{\Psi} \right\}. \end{aligned} \quad (1.17)$$

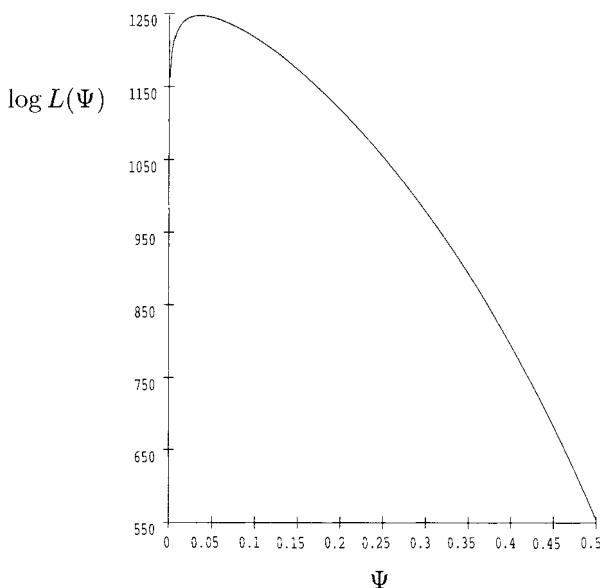


Figure 1.1 Plot of log likelihood function $\log L(\Psi)$ for the multinomial data in Thisted (1988).

Table 1.1 Results of the Newton–Raphson and Scoring Algorithms for the Example 1.1 for Data in Thisted (1988).

Iteration	Newton–Raphson		Scoring	
	$\Psi^{(k)}$	$S(\mathbf{y}; \Psi^{(k)})$	$\Psi^{(k)}$	$S(\mathbf{y}; \Psi^{(k)})$
0	0.05704611	−387.74068038	0.05704611	−387.74068038
1	0.02562679	376.95646890	0.03698326	−33.88267279
2	0.03300085	80.19367817	0.03579085	−2.15720180
3	0.03552250	5.24850707	0.03571717	−0.13386352
4	0.03571138	0.02527096	0.03571260	−0.00829335
5	0.03571230	0.00000059	0.03571232	−0.00051375
6	0.03571230	−0.00000000	0.03571230	−0.00003183

Source: Adapted from Thisted (1988).

Suppose now that the first of the original four multinomial cells, which has an associated probability of $\frac{1}{2} + \frac{1}{4}\Psi$, could be split into two subcells having probabilities $\frac{1}{2}$ and $\frac{1}{4}\Psi$, respectively, and let y_{11} and y_{12} be the corresponding split of y_1 , where

$$y_1 = y_{11} + y_{12}.$$

Then on modifying the likelihood equation (1.14) according to this split, it is clear that the MLE of Ψ on the basis of this split is simply

$$(y_{12} + y_4)/(y_{12} + y_2 + y_3 + y_4). \quad (1.18)$$

This is because in effect the modified likelihood function for Ψ has the same form as that obtained by considering $y_{12} + y_4$ to be a realization from a binomial distribution with sample size $y_{12} + y_2 + y_3 + y_4$ and probability parameter Ψ .

We now formalize this approach through the application of the EM algorithm. The observed vector of frequencies \mathbf{y} is viewed as being incomplete and the complete-data vector is taken to be

$$\mathbf{x} = (y_{11}, y_{12}, y_2, y_3, y_4)^T.$$

The cell frequencies in \mathbf{x} are assumed to arise from a multinomial distribution having five cells with probabilities

$$\frac{1}{2}, \frac{1}{4}\Psi, \frac{1}{4}(1 - \Psi), \frac{1}{4}(1 - \Psi), \text{ and } \frac{1}{4}\Psi. \quad (1.19)$$

In this framework, y_{11} and y_{12} are regarded as the unobservable or missing data, since we only get to observe their sum y_1 .

If we take the distribution of the complete-data vector \mathbf{X} to be multinomial with n draws with respect to now five cells with probabilities specified by (1.19), then it implies that the observable or incomplete-data vector \mathbf{Y} has its original multinomial distribution with cell probabilities specified by (1.11). This can be confirmed by verifying that

$$g(\mathbf{y}; \Psi) = \sum g_c(\mathbf{x}; \Psi), \quad (1.20)$$

where

$$g_c(\mathbf{x}; \Psi) = C(\mathbf{x})(\frac{1}{2})^{y_{11}}(\frac{1}{4}\Psi)^{y_{12}}(\frac{1}{4} - \frac{1}{4}\Psi)^{y_2}(\frac{1}{4} - \frac{1}{4}\Psi)^{y_3}(\frac{1}{4}\Psi)^{y_4} \quad (1.21)$$

and

$$C(\mathbf{x}) = \frac{n!}{y_{11}!y_{12}!y_2!y_3!y_4!},$$

and where the summation in (1.20) is over all values of \mathbf{x} such that

$$y_{11} + y_{12} = y_1.$$

From (1.21), the complete-data log likelihood is, therefore, apart from a term not involving Ψ ,

$$\log L_c(\Psi) = (y_{12} + y_4) \log \Psi + (y_2 + y_3) \log(1 - \Psi). \quad (1.22)$$

On equating the derivative of (1.22) with respect to Ψ to zero and solving for Ψ , we find that the complete-data MLE of Ψ is given by (1.18).

Since the frequency y_{12} is unobservable, we are unable to estimate Ψ by (1.18). With the EM algorithm, this obstacle is overcome by the E-step, as it handles the problem of filling in for unobservable data by averaging the complete-data log likelihood over its conditional distribution given the observed data \mathbf{y} . But in order to calculate this conditional expectation, we have to specify a value for Ψ . Let $\Psi^{(0)}$ be the value specified initially for Ψ . Then on the first iteration of the EM algorithm, the E-step requires the computation of the conditional expectation of $\log L_c(\Psi)$ given \mathbf{y} , using $\Psi^{(0)}$ for Ψ , which can be written as

$$Q(\Psi; \Psi^{(0)}) = E_{\Psi^{(0)}} \{\log L_c(\Psi) \mid \mathbf{y}\}.$$

As $\log L_c(\Psi)$ is a linear function of the unobservable data y_{11} and y_{12} for this problem, the E-step is effected simply by replacing y_{11} and y_{12} by their current conditional expectations given the observed data y . Considering the random variable Y_{11} , corresponding to y_{11} , it is easy to verify that conditional on y , effectively y_1 , Y_{11} has a binomial distribution with sample size y_1 and probability parameter

$$\frac{1}{2}/\left(\frac{1}{2} + \frac{1}{4}\Psi^{(0)}\right),$$

where $\Psi^{(0)}$ is used in place of the unknown parameter Ψ . Thus the initial conditional expectation of Y_{11} given y_1 is

$$E_{\Psi^{(0)}}(Y_{11} | y_1) = y_{11}^{(0)},$$

where

$$y_{11}^{(0)} = \frac{1}{2}y_1/\left(\frac{1}{2} + \frac{1}{4}\Psi^{(0)}\right). \quad (1.23)$$

This completes the E-step on the first iteration since

$$\begin{aligned} y_{12}^{(0)} &= y_1 - y_{11}^{(0)} \\ &= \frac{1}{4}y_1\Psi^{(0)}/\left(\frac{1}{2} + \frac{1}{4}\Psi^{(0)}\right). \end{aligned} \quad (1.24)$$

The M-step is undertaken on the first iteration by choosing $\Psi^{(1)}$ to be the value of Ψ that maximizes $Q(\Psi; \Psi^{(0)})$ with respect to Ψ . Since this Q -function is given simply by replacing the unobservable frequencies y_{11} and y_{12} with their current conditional expectations $y_{11}^{(0)}$ and $y_{12}^{(0)}$ in the complete-data log likelihood, $\Psi^{(1)}$ is obtained by substituting $y_{12}^{(0)}$ for y_{12} in (1.18) to give

$$\begin{aligned} \Psi^{(1)} &= (y_{12}^{(0)} + y_4)/(y_{12}^{(0)} + y_2 + y_3 + y_4) \\ &= (y_{12}^{(0)} + y_4)/(n - y_{11}^{(0)}). \end{aligned} \quad (1.25)$$

This new fit $\Psi^{(1)}$ for Ψ is then substituted for Ψ into the right-hand sides of (1.23) and (1.24) to produce updated values $y_{11}^{(1)}$ and $y_{12}^{(1)}$ for y_{11} and y_{12} for use in place of $y_{11}^{(0)}$ and $y_{12}^{(0)}$ in the right-hand side of (1.25). This leads to a new fit $\Psi^{(2)}$ for Ψ , and so on. It follows on so alternating the E- and M-steps on the $(k+1)$ th iteration of the EM algorithm that

$$\Psi^{(k+1)} = (y_{12}^{(k)} + y_4)/(n - y_{11}^{(k)}) \quad (1.26)$$

where

$$y_{11}^{(k)} = \frac{1}{2}y_1/\left(\frac{1}{2} + \frac{1}{4}\Psi^{(k)}\right)$$

and

$$y_{12}^{(k)} = y_1 - y_{11}^{(k)}.$$

On putting

$$\Psi^{(k+1)} = \Psi^{(k)} = \Psi^*$$

in (1.26), we can explicitly solve the resulting quadratic equation in Ψ^* to confirm that the sequence of EM iterates $\{\Psi^{(k)}\}$, irrespective of the starting value $\Psi^{(0)}$, converges to the MLE of Ψ obtained by directly solving the (incomplete-data) likelihood equation given by equating (1.15) to zero.

Table 1.2 Results of EM Algorithm for Example 1.1 for Data in DLR .

Iteration	$\Psi^{(k)}$	$\Psi^{(k)} - \hat{\Psi}$	$r^{(k)}$	$\log L(\Psi^{(k)})$
0	0.500000000	0.126821498	—	64.62974
1	0.608247423	0.018574075	0.1465	67.32017
2	0.624321051	0.002500447	0.1346	67.38292
3	0.626488879	0.000332619	0.1330	67.38408
4	0.626777323	0.000044176	0.1328	67.38410
5	0.626815632	0.000005866	0.1328	67.38410
6	0.626820719	0.000000779	0.1328	67.38410
7	0.626821395	0.000000104	0.1328	67.38410
8	0.626821484	0.000000014	—	67.38410

Source: Adapted from Dempster et al. (1977).

In Tables 1.2 and 1.3, we report the results of the EM algorithm applied to the data sets considered by DLR and Thisted (1988), respectively. In Table 1.2, we see that starting from an initial value of $\Psi^{(0)} = 0.5$, the EM algorithm moved for eight iterations. The third column in this table gives the deviation $\Psi^{(k)} - \hat{\Psi}$, and the fourth column gives the ratio of successive deviations

$$r^{(k)} = (\Psi^{(k+1)} - \Psi^{(k)}) / (\Psi^{(k)} - \Psi^{(k-1)}).$$

It can be seen that $r^{(k)}$ is essentially constant for $k \geq 3$ consistent with a linear rate of convergence equal to 0.1328. This rate of convergence is to be established in general for the EM algorithm in Section 3.9.

On comparing the results of the EM algorithm in Table 1.3 for the data set in Thisted (1988) with those of the Newton-Raphson and scoring methods in Table 1.1, we see that the EM algorithm takes about 15 more iterations to converge to the MLE.

1.4.3 Example 1.2: Estimation of Mixing Proportions

As to be discussed further in Section 1.8, the publication of the DLR paper greatly stimulated interest in the use of finite mixture distributions to model heterogeneous data. This is because the fitting of mixture models by maximum likelihood is a classic example of a problem that is simplified considerably by the EM's conceptual unification of ML estimation from data that can be viewed as being incomplete.

A wide variety of applications of finite mixture models are given in McLachlan and Basford (1988) and McLachlan and Peel (2000a). We consider here an example involving the estimation of the proportions in which the components of the mixture occur, where the component densities are completely specified. In the next chapter, we consider the more difficult case where the component densities are specified up to a number of unknown parameters that have to be estimated along with the mixing proportions. But the former case of completely specified component densities is not unrealistic, as there are situations in practice where there are available separate samples from each of the component distributions of the mixture that enable the component densities to be estimated with adequate precision before the fitting of the mixture model; see McLachlan (1992, Chapter 2).

Table 1.3 Results of EM Algorithm for Example 1.1 for Data in Thisted (1988).

Iteration	$\Psi^{(k)}$	$\Psi^{(k)} - \hat{\Psi}$	$r^{(k)}$	$\log L(\Psi^{(k)})$
0	0.05704611	0.0213338	—	1242.4180
1	0.04600534	0.0103411	0.48473	1245.8461
2	0.04077975	0.0005067	0.49003	1246.7791
3	0.03820858	0.0002496	0.49261	1247.0227
4	0.03694516	0.0001233	0.49388	1247.0845
5	0.03632196	0.0000610	0.49450	1247.0999
6	0.03605397	0.0000302	0.49481	1247.1047
7	0.03586162	0.0000149	0.49496	1247.1049
8	0.03578622	0.0000074	0.49502	1247.1050
9	0.03574890	0.0000037	0.49503	1247.1050
10	0.03573042	0.0000018	0.49503	1247.1050
11	0.03572127	0.0000009	0.49503	1247.1050
12	0.03571674	0.0000004	0.49503	1247.1050
13	0.03571450	0.0000002	0.49503	1247.1050
14	0.03571339	0.0000001	0.49503	1247.1050
15	0.03571284	0.0000001	0.49503	1247.1050
16	0.03571257	0.0000000	0.49503	1247.1050
17	0.03571243	0.0000000	0.49503	1247.1050
18	0.03571237	0.0000000	0.49503	1247.1050
19	0.03571233	0.0000000	0.49503	1247.1050
20	0.03571232	0.0000000	0.49503	1247.1050
21	0.03571231	0.0000000	0.49503	1247.1050
22	0.03571231	0.0000000	—	1247.1050

Suppose that the p.d.f. of a random vector \mathbf{W} has a g -component mixture form

$$f(\mathbf{w}; \boldsymbol{\Psi}) = \sum_{i=1}^g \pi_i f_i(\mathbf{w}), \quad (1.27)$$

where $\boldsymbol{\Psi} = (\pi_1, \dots, \pi_{g-1})^T$ is the vector containing the unknown parameters, namely the $g-1$ mixing proportions π_1, \dots, π_{g-1} , since

$$\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i.$$

The component p.d.f.'s $f_i(\mathbf{w})$ are completely specified.

This mixture model covers situations where the underlying population is modeled as consisting of g distinct groups G_1, \dots, G_g in some unknown proportions π_1, \dots, π_g , and where the conditional p.d.f. of \mathbf{W} given membership of the i th group G_i is $f_i(\mathbf{w})$. For example, in the problem considered by Do and McLachlan (1984), the population of interest consists of rats from g species G_1, \dots, G_g , that are consumed by owls in some unknown proportions π_1, \dots, π_g . The problem is to estimate the π_i on the basis of the observation vector \mathbf{W} containing measurements recorded on a sample of size n of rat skulls

taken from owl pellets. The rats constitute part of an owl's diet, and indigestible material is regurgitated as a pellet.

We let

$$\mathbf{y} = (\mathbf{w}_1, \dots, \mathbf{w}_n)^T$$

denote the observed random sample obtained from the mixture density (1.27). The log likelihood function for Ψ that can be formed from the observed data \mathbf{y} is given by

$$\begin{aligned}\log L(\Psi) &= \sum_{j=1}^n \log f(\mathbf{w}_j; \Psi) \\ &= \sum_{j=1}^n \log \left\{ \sum_{i=1}^g \pi_i f_i(\mathbf{w}_j) \right\}. \end{aligned} \quad (1.28)$$

On differentiating (1.28) with respect to π_i ($i = 1, \dots, g - 1$) and equating the result to zero, we obtain

$$\sum_{j=1}^n \left\{ \frac{f_i(\mathbf{w}_j)}{f(\mathbf{w}_j; \hat{\Psi})} - \frac{f_g(\mathbf{w}_j)}{f(\mathbf{w}_j; \hat{\Psi})} \right\} = 0 \quad (i = 1, \dots, g - 1) \quad (1.29)$$

as the likelihood equation, which clearly does not yield an explicit solution for

$$\hat{\Psi} = (\hat{\pi}_1, \dots, \hat{\pi}_{g-1})^T.$$

In order to pose this problem as an incomplete-data one, we now introduce as the unobservable or missing data the vector

$$\mathbf{z} = (z_1^T, \dots, z_n^T)^T, \quad (1.30)$$

where \mathbf{z}_j is a g -dimensional vector of zero-one indicator variables and where $z_{ij} = (z_j)_i$ is one or zero according to whether \mathbf{w}_j arose or did not arise from the i th component of the mixture ($i = 1, \dots, g$; $j = 1, \dots, n$). Of course in some applications (such as in the rat data one above), the components of the mixture correspond to externally existing groups and so each realization \mathbf{w}_j in the observed random sample from the mixture density does have a tangible component membership. But in other applications, component membership of the realizations is just a conceptual device to formulate the problem within the EM framework.

If these z_{ij} were observable, then the MLE of π_i is simply given by

$$\sum_{j=1}^n z_{ij}/n \quad (i = 1, \dots, g), \quad (1.31)$$

which is the proportion of the sample having arisen from the i th component of the mixture.

As in the last example, the EM algorithm handles the addition of the unobservable data to the problem by working with the current conditional expectation of the complete-data log likelihood given the observed data. On defining the complete-data vector \mathbf{x} as

$$\mathbf{x} = (\mathbf{y}^T, \mathbf{z}^T)^T, \quad (1.32)$$

the complete-data log likelihood for Ψ has the multinomial form

$$\log L_c(\Psi) = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log \pi_i + C, \quad (1.33)$$

where

$$C = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log f_i(\mathbf{w}_j)$$

does not depend on Ψ .

As (1.33) is linear in the unobservable data z_{ij} , the E-step (on the $(k+1)$ th iteration) simply requires the calculation of the current conditional expectation of Z_{ij} given the observed data \mathbf{y} , where Z_{ij} is the random variable corresponding to z_{ij} . Now

$$\begin{aligned} E_{\Psi^{(k)}}(Z_{ij} | \mathbf{y}) &= \text{pr}_{\Psi^{(k)}}\{Z_{ij} = 1 | \mathbf{y}\} \\ &= z_{ij}^{(k)}, \end{aligned} \quad (1.34)$$

where by Bayes Theorem,

$$z_{ij}^{(k)} = \tau_i(\mathbf{w}_j; \Psi^{(k)}) \quad (1.35)$$

and

$$\tau_i(\mathbf{w}_j; \Psi^{(k)}) = \pi_i^{(k)} f_i(\mathbf{w}_j) / f(\mathbf{w}_j; \Psi^{(k)}) \quad (1.36)$$

for $i = 1, \dots, g$; $j = 1, \dots, n$. The quantity $\tau_i(\mathbf{w}_j; \Psi^{(k)})$ is the posterior probability that the j th member of the sample with observed value \mathbf{w}_j belongs to the i th component of the mixture.

The M-step on the $(k+1)$ th iteration simply requires replacing each z_{ij} by $z_{ij}^{(k)}$ in (1.31) to give

$$\pi_i^{(k+1)} = \sum_{j=1}^n z_{ij}^{(k)} / n \quad (1.37)$$

for $i = 1, \dots, g$. Thus in forming the estimate of π_i on the $(k+1)$ th iteration, there is a contribution from each observation \mathbf{w}_j equal to its (currently assessed) posterior probability of membership of the i th component of the mixture model. This EM solution therefore has an intuitively appealing interpretation.

The computation of the MLE of π_i by direct maximization of the incomplete-data log likelihood function (1.28) requires solving the likelihood equation (1.29). The latter can be identified with the iterative solution (1.37) provided by the EM algorithm after some manipulation as follows. On multiplying throughout by $\hat{\pi}_i$ in equation (1.29), we have that

$$\sum_{j=1}^n \{\tau_i(\mathbf{w}_j; \hat{\Psi}) - (\hat{\pi}_i / \hat{\pi}_g) \tau_g(\mathbf{w}_j; \hat{\Psi})\} = 0 \quad (i = 1, \dots, g-1). \quad (1.38)$$

As (1.38) also holds for $i = g$, we can sum over $i = 1, \dots, g$ in (1.38) to give

$$\hat{\pi}_g = \sum_{j=1}^n \tau_g(\mathbf{w}_j; \hat{\Psi}) / n. \quad (1.39)$$

Substitution now of (1.39) into (1.38) yields

$$\hat{\pi}_i = \sum_{j=1}^n \tau_i(\mathbf{w}_j; \hat{\Psi}) / n \quad (1.40)$$

for $i = 1, \dots, g-1$, which also holds for $i = g$ from (1.39). The resulting equation (1.40) for the MLE $\hat{\pi}_i$ can be identified with the iterative solution (1.36) given by the EM

algorithm. The latter solves the likelihood equation by substituting an initial value for Ψ into the right-hand side of (1.40), which yields a new estimate for Ψ , which in turn is substituted into the right-hand side of (1.40) to yield a new estimate, and so on until convergence.

Even before the formulation of the EM algorithm by DLR, various researchers have carried out these manipulations in their efforts to solve the likelihood equation for mixture models with specific component densities; see, for example, Hasselblad (1966, 1969), Wolfe (1967, 1970), and Day (1969). As demonstrated above for the estimation of the mixing proportions, the application of the EM algorithm to the mixture problem automatically reveals the iterative scheme to be followed for the computation of the MLE. Furthermore, it ensures that the likelihood values increase monotonically. Prior to DLR, various researchers did note the monotone convergence of the likelihood sequences produced in their particular applications, but were only able to speculate on this monotonicity holding in general.

As the $z_{ij}^{(k)}$ are probabilities, the E-step of the EM algorithm effectively imputes fractional values for the unobservable zero-one indicator variables z_{ij} . In so doing it avoids the biases associated with *ad hoc* iterative procedures that impute only zero-one values; that is, that insist on outright component membership for each observation at each stage. For example, for each j ($j = 1, \dots, n$), let

$$\hat{z}_{ij}^{(k)} = 1 \text{ if } i = \arg \max_h \tau_h(\mathbf{w}_j; \Psi^{(k)}),$$

and zero otherwise; this is equivalent to assigning the j th observation \mathbf{w}_j to the component of the mixture for which it has the highest (currently assessed) posterior probability of belonging. If these zero-one values are imputed for the z_{ij} in updating the estimate of π_i from (1.31), then this in general will produce biased estimates of the mixing proportions (McLachlan and Basford, 1988, Chapter 4).

Numerical Example. As a numerical example, we generated a random sample of $n = 50$ observations w_1, \dots, w_n from a mixture of two univariate normal densities with means $\mu_1 = 0$ and $\mu_2 = 2$ and common variance $\sigma^2 = 1$ in proportions $\pi_1 = 0.8$ and $\pi_2 = 0.2$. Starting the EM algorithm from $\pi_1^{(0)} = 0.5$, it converged after 27 iterations to the solution $\hat{\pi}_1 = 0.75743$. The EM algorithm was stopped when

$$|\pi_1^{(k+1)} - \pi_1^{(k)}| < 10^{-5}.$$

It was also started from the moment estimate given by

$$\begin{aligned}\tilde{\pi}_1 &= (\bar{w} - \mu_2)/(\mu_1 - \mu_2) \\ &= 0.86815\end{aligned}$$

and, using the same stopping criterion, it converged after 30 iterations to $\hat{\pi}_1$.

In Table 1.4, we have listed the value of $\pi_1^{(k)}$ and of $\log L(\pi_1^{(k)})$ for various values of k . It can be seen that it is during the first few iterations that the EM algorithm makes most of its progress in reaching the maximum value of the log likelihood function.

Table 1.4 Results of EM Algorithm for Example on Estimation of Mixing Proportions.

Iteration <i>k</i>	$\pi_1^{(k)}$	$\log L(\pi_1^{(k)})$
0	0.50000	-91.87811
1	0.68421	-85.55353
2	0.70304	-85.09035
3	0.71792	-84.81398
4	0.72885	-84.68609
5	0.73665	-84.63291
6	0.74218	-84.60978
7	0.74615	-84.58562
:	:	:
27	0.75743	-84.58562

1.5 FORMULATION OF THE EM ALGORITHM

1.5.1 EM Algorithm

We let \mathbf{Y} be the random vector corresponding to the observed data \mathbf{y} , having p.d.f. postulated as $g(\mathbf{y}; \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\Psi_1, \dots, \Psi_d)^T$ is a vector of unknown parameters with parameter space Ω .

The EM algorithm is a broadly applicable algorithm that provides an iterative procedure for computing MLE's in situations where, but for the absence of some additional data, ML estimation would be straightforward. Hence in this context, the observed data vector \mathbf{y} is viewed as being incomplete and is regarded as an observable function of the so-called complete data. The notion of 'incomplete data' includes the conventional sense of missing data, but it also applies to situations where the complete data represent what would be available from some hypothetical experiment. In the latter case, the complete data may contain some variables that are never observable in a data sense. Within this framework, we let \mathbf{x} denote the vector containing the augmented or so-called complete data, and we let \mathbf{z} denote the vector containing the additional data, referred to as the unobservable or missing data.

As will become evident from the many examples of the EM algorithm discussed in this book, even when a problem does not at first appear to be an incomplete-data one, computation of the MLE is often greatly facilitated by artificially formulating it to be as such. This is because the EM algorithm exploits the reduced complexity of ML estimation given the complete data. For many statistical problems the complete-data likelihood has a nice form.

We let $g_c(\mathbf{x}; \boldsymbol{\Psi})$ denote the p.d.f. of the random vector \mathbf{X} corresponding to the complete-data vector \mathbf{x} . Then the complete-data log likelihood function that could be formed for $\boldsymbol{\Psi}$ if \mathbf{x} were fully observable is given by

$$\log L_c(\boldsymbol{\Psi}) = \log g_c(\mathbf{x}; \boldsymbol{\Psi}).$$

Formally, we have two sample spaces \mathcal{X} and \mathcal{Y} and a many-to-one mapping from \mathcal{X} to \mathcal{Y} . Instead of observing the complete-data vector \mathbf{x} in \mathcal{X} , we observe the incomplete-data

vector $\mathbf{y} = \mathbf{y}(\mathbf{x})$ in \mathcal{Y} . It follows that

$$g(\mathbf{y}; \boldsymbol{\Psi}) = \int_{\mathcal{X}(\mathbf{y})} g_c(\mathbf{x}; \boldsymbol{\Psi}) d\mathbf{x},$$

where $\mathcal{X}(\mathbf{y})$ is the subset of \mathcal{X} determined by the equation $\mathbf{y} = \mathbf{y}(\mathbf{x})$.

The EM algorithm approaches the problem of solving the incomplete-data likelihood equation (1.1) indirectly by proceeding iteratively in terms of the complete-data log likelihood function, $\log L_c(\boldsymbol{\Psi})$. As it is unobservable, it is replaced by its conditional expectation given \mathbf{y} , using the current fit for $\boldsymbol{\Psi}$.

More specifically, let $\boldsymbol{\Psi}^{(0)}$ be some initial value for $\boldsymbol{\Psi}$. Then on the first iteration, the E-step requires the calculation of

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)}) = E_{\boldsymbol{\Psi}^{(0)}} \{ \log L_c(\boldsymbol{\Psi}) \mid \mathbf{y} \}.$$

The M-step requires the maximization of $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(0)})$ with respect to $\boldsymbol{\Psi}$ over the parameter space Ω . That is, we choose $\boldsymbol{\Psi}^{(1)}$ such that

$$Q(\boldsymbol{\Psi}^{(1)}; \boldsymbol{\Psi}^{(0)}) \geq Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)})$$

for all $\boldsymbol{\Psi} \in \Omega$. The E- and M-steps are then carried out again, but this time with $\boldsymbol{\Psi}^{(0)}$ replaced by the current fit $\boldsymbol{\Psi}^{(1)}$. On the $(k+1)$ th iteration, the E- and M-steps are defined as follows:

E-Step. Calculate $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$, where

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}} \{ \log L_c(\boldsymbol{\Psi}) \mid \mathbf{y} \}. \quad (1.41)$$

M-Step. Choose $\boldsymbol{\Psi}^{(k+1)}$ to be any value of $\boldsymbol{\Psi} \in \Omega$ that maximizes $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$; that is,

$$Q(\boldsymbol{\Psi}^{(k+1)}; \boldsymbol{\Psi}^{(k)}) \geq Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) \quad (1.42)$$

for all $\boldsymbol{\Psi} \in \Omega$.

The E- and M-steps are alternated repeatedly until the difference

$$L(\boldsymbol{\Psi}^{(k+1)}) - L(\boldsymbol{\Psi}^{(k)})$$

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\boldsymbol{\Psi}^{(k)})\}$. DLR show that the (incomplete-data) likelihood function $L(\boldsymbol{\Psi})$ is not decreased after an EM iteration; that is,

$$L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k)}) \quad (1.43)$$

for $k = 0, 1, 2, \dots$. Hence convergence must be obtained with a sequence of likelihood values that are bounded above.

Another way of expressing (1.42) is to say that $\boldsymbol{\Psi}^{(k+1)}$ belongs to

$$\mathcal{M}(\boldsymbol{\Psi}^{(k)}) = \arg \max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}), \quad (1.44)$$

which is the set of points that maximize $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$.

We see from the above that it is not necessary to specify the exact mapping from \mathcal{X} to \mathcal{Y} , nor the corresponding representation of the incomplete-data density g in terms of the complete-data density g_c . All that is necessary is the specification of the complete-data vector \boldsymbol{x} and the conditional density of \boldsymbol{X} given the observed data vector \boldsymbol{y} . Specification of this conditional density is needed in order to carry out the E-step. As the choice of the complete-data vector \boldsymbol{x} is not unique, it is chosen for computational convenience with respect to carrying out the E- and M-steps. Consideration has been given to the choice of \boldsymbol{x} so as to speed up the convergence of the corresponding EM algorithm; see Section 5.12.

As pointed out by a referee of the DLR paper, the use of the term “algorithm” to describe this procedure can be criticized, “because it does not specify the sequence of steps actually required to carry out a single E- or M-step.” The EM algorithm is really a generic device. Hunter (2003) goes so far as to suggest the usage “EM algorithms” or “an EM algorithm” because many different examples fall under the EM umbrella.

1.5.2 Example 1.3: Censored Exponentially Distributed Survival Times

We suppose W is a nonnegative random variable having an exponential distribution with mean μ . Thus its probability density function (p.d.f.) is given by

$$f(w; \mu) = \mu^{-1} \exp(-w/\mu) I_{(0, \infty)}(w) \quad (\mu > 0), \quad (1.45)$$

where the indicator function $I_{(0, \infty)}(w) = 1$ for $w > 0$ and is zero elsewhere. The distribution function is given by

$$F(w; \mu) = \{1 - \exp(-w/\mu)\}I_{(0, \infty)}(w).$$

In survival or reliability analyses, a study to observe a random sample W_1, \dots, W_n from (1.45) will generally be terminated in practice before all of these random variables are able to be observed. We let

$$\boldsymbol{y} = (\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_n^T)^T$$

denote the observed data, where

$$\boldsymbol{y}_j = (c_j, \delta_j)^T$$

and $\delta_j = 0$ or 1 according as the observation W_j is censored or uncensored at c_j ($j = 1, \dots, n$). That is, if the observation W_j is uncensored, its realized value w_j is equal to c_j , whereas if it is censored at c_j , then w_j is some value greater than c_j ($j = 1, \dots, n$).

In this example, the unknown parameter vector Ψ is a scalar, being equal to μ . We suppose now that the observations have been relabeled so that W_1, \dots, W_r denote the r uncensored observations and W_{r+1}, \dots, W_n the $n - r$ censored observations. The log likelihood function for μ formed on the basis of \boldsymbol{y} is given by

$$\log L(\mu) = -r \log \mu - \sum_{j=1}^n c_j / \mu. \quad (1.46)$$

In this case, the MLE of μ can be derived explicitly from equating the derivative of (1.46) to zero to give

$$\hat{\mu} = \sum_{j=1}^n c_j / r. \quad (1.47)$$

Thus there is no need for the iterative computation of $\hat{\mu}$. But in this simple case, it is instructive to demonstrate how the EM algorithm would work.

The complete-data vector \mathbf{x} can be declared to be

$$\begin{aligned}\mathbf{x} &= (w_1, \dots, w_n)^T \\ &= (w_1, \dots, w_r, \mathbf{z}^T)^T,\end{aligned}$$

where

$$\mathbf{z} = (w_{r+1}, \dots, w_n)^T$$

contains the unobservable realizations of the $n - r$ censored random variables. In this example, the so-called unobservable or missing vector \mathbf{z} is potentially observable in a data sense, as if the experiment were continued until each item failed, then there would be no censored observations.

The complete-data log likelihood is given by

$$\begin{aligned}\log L_c(\mu) &= \sum_{j=1}^n \log g_c(w_j; \mu) \\ &= -n \log \mu - \mu^{-1} \sum_{j=1}^n w_j.\end{aligned}\tag{1.48}$$

It can be seen that $L_c(\mu)$ belongs to the regular exponential family. We shall proceed now without making explicit use of this property, but in the next section, we shall show how it can be exploited to simplify the implementation of the EM algorithm.

As $L_c(\mu)$ can be seen to be linear in the unobservable data w_{r+1}, \dots, w_n , the calculation of $Q(\mu; \mu^{(k)})$ on the E-step (on the $(k + 1)$ th iteration) simply requires each such w_j to be replaced by its conditional expectation given the observed data \mathbf{y} , using the current fit $\mu^{(k)}$ for μ . By the lack of memory of the exponential distribution, the conditional distribution of $W_j - c_j$ given that $W_j > c_j$ is still exponential with mean μ . Equivalently, the conditional p.d.f of W_j given that it is greater than c_j is

$$\mu^{-1} \exp\{-(w_j - c_j)/\mu\} I_{(c_j, \infty)}(w_j) \quad (\mu > 0).\tag{1.49}$$

From (1.49), we have that

$$\begin{aligned}E_{\mu^{(k)}}(W_j | \mathbf{y}) &= E_{\mu^{(k)}}(W_j | W_j > c_j) \\ &= c_j + E_{\mu^{(k)}}(W_j) \\ &= c_j + \mu^{(k)}\end{aligned}\tag{1.50}$$

for $j = r + 1, \dots, n$.

On using (1.50) to take the current conditional expectation of the complete-data log likelihood $\log L_c(\mu)$, we have that

$$\begin{aligned}Q(\mu; \mu^{(k)}) &= -n \log \mu - \mu^{-1} \left\{ \sum_{j=1}^r c_j + \sum_{j=r+1}^n (c_j + \mu^{(k)}) \right\} \\ &= -n \log \mu - \mu^{-1} \left\{ \sum_{j=1}^n c_j + (n - r)\mu^{(k)} \right\}.\end{aligned}\tag{1.51}$$

Concerning the M-step on the $(k + 1)$ th iteration, it follows from (1.51) that the value of μ that maximizes $Q(\mu; \mu^{(k)})$ is given by the MLE of μ that would be formed from the

complete data, but with each unobservable w_j replaced by its current conditional expectation given by (1.50). Accordingly,

$$\begin{aligned}\mu^{(k+1)} &= \left\{ \sum_{j=1}^r c_j + \sum_{j=r+1}^n E_{\mu^{(k)}}(W_j \mid \mathbf{y}) \right\} / n \\ &= \left\{ \sum_{j=1}^r c_j + \sum_{j=r+1}^n (c_j + \mu^{(k)}) \right\} / n \\ &= \left\{ \sum_{j=1}^n c_j + (n - r)\mu^{(k)} \right\} / n.\end{aligned}\quad (1.52)$$

On putting $\mu^{(k+1)} = \mu^{(k)} = \mu^*$ in (1.52) and solving for μ^* , we have for $r < n$ that $\mu^* = \hat{\mu}$. That is, the EM sequence $\{\mu^{(k)}\}$ has the MLE $\hat{\mu}$ as its unique limit point, as $k \rightarrow \infty$.

In order to demonstrate the rate of convergence of this sequence to $\hat{\mu}$, we can from (1.52) express $\mu^{(k+1)}$ in terms of the MLE $\hat{\mu}$ as

$$\begin{aligned}\mu^{(k+1)} &= \{r\hat{\mu} + (n - r)\mu^{(k)}\} / n \\ &= \hat{\mu} + n^{-1}(n - r)(\mu^{(k)} - \hat{\mu}),\end{aligned}$$

which gives

$$\mu^{(k+1)} - \hat{\mu} = (1 - r/n)(\mu^{(k)} - \hat{\mu}). \quad (1.53)$$

This establishes that $\mu^{(k)}$ converges to $\hat{\mu}$, as $k \rightarrow \infty$, provided $r < n$.

It can be seen for this problem that each EM iteration is linear. We shall see later that in general the rate of convergence of the EM algorithm is essentially linear. The rate of convergence here is $(1 - r/n)$, which is the proportion of censored observations in the observed sample. This proportion can be viewed as the missing information in the sample, as will be made more precise in Section 3.9.

1.5.3 E- and M-Steps for the Regular Exponential Family

The complete-data p.d.f. $g_c(\mathbf{x}; \boldsymbol{\Psi})$ is from an exponential family if

$$g_c(\mathbf{x}; \boldsymbol{\Psi}) = \exp\{\mathbf{a}^T(\boldsymbol{\Psi})\mathbf{t}(\mathbf{x}) - b(\boldsymbol{\Psi}) + c(\mathbf{x})\}, \quad (1.54)$$

where the sufficient statistic $\mathbf{t}(\mathbf{x})$ is a $k \times 1$ ($k \geq d$) vector and $\mathbf{a}(\boldsymbol{\Psi})$ is a $k \times 1$ vector function of the $d \times 1$ parameter vector $\boldsymbol{\Psi}$, and $b(\boldsymbol{\Psi})$ and $c(\mathbf{x})$ are scalar functions. The parameter space Ω is a d -dimensional convex set such that (1.54) defines a p.d.f. for all $\boldsymbol{\Psi}$ in Ω ; that is,

$$\Omega = \{\boldsymbol{\Psi}: \int_{\mathcal{X}} \exp\{\mathbf{a}^T(\boldsymbol{\Psi})\mathbf{t}(\mathbf{x}) + c(\mathbf{x})\} d\mathbf{x} < \infty\}. \quad (1.55)$$

If $k = d$ and the Jacobian of $\mathbf{a}(\boldsymbol{\Psi})$ is of full rank, then $g_c(\mathbf{x}; \boldsymbol{\Psi})$ is said to be from a regular exponential family. The coefficient $\mathbf{a}(\boldsymbol{\Psi})$ of the sufficient statistic $\mathbf{t}(\mathbf{x})$ in (1.54) is referred to as the natural or canonical parameter (vector). Thus if the complete-data p.d.f. $g_c(\mathbf{x}; \boldsymbol{\Psi})$ is from a regular exponential family in canonical form, then

$$g_c(\mathbf{x}; \boldsymbol{\Psi}) = \exp\{\boldsymbol{\Psi}^T \mathbf{t}(\mathbf{x}) - b(\boldsymbol{\Psi}) + c(\mathbf{x})\}. \quad (1.56)$$

The parameter Ψ in (1.56) is unique up to an arbitrary nonsingular $d \times d$ linear transformation, as is the corresponding choice of $t(\mathbf{X})$.

The expectation of the sufficient statistic $t(\mathbf{X})$ in (1.56) is given by

$$E_{\Psi}\{t(\mathbf{X})\} = \partial b(\Psi)/\partial\Psi. \quad (1.57)$$

Another property of the regular exponential family, which we shall use in a later section, is that the expected information matrix for the natural parameter vector equals the covariance matrix of the sufficient statistic $t(\mathbf{X})$. Thus we have for the regular exponential family in the canonical form (1.56) that

$$\text{cov}_{\Psi}\{t(\mathbf{X})\} = \mathcal{I}_c(\Psi), \quad (1.58)$$

where since the second derivatives of (1.56) do not depend on the data,

$$\begin{aligned} \mathcal{I}_c(\Psi) &= -\partial^2 \log L_c(\Psi)/\partial\Psi\partial\Psi^T \\ &= \partial^2 b(\Psi)/\partial\Psi\partial\Psi^T. \end{aligned} \quad (1.59)$$

On taking the conditional expectation of $\log L_c(\Psi)$ given \mathbf{y} , we have from (1.56) that $Q(\Psi; \Psi^{(k)})$ is given by, ignoring terms not involving Ψ ,

$$Q(\Psi; \Psi^{(k)}) = \Psi^T t^{(k)} - b(\Psi), \quad (1.60)$$

where

$$t^{(k)} = E_{\Psi^{(k)}}\{t(\mathbf{X}) \mid \mathbf{y}\}$$

and where $\Psi^{(k)}$ denotes the current fit for Ψ .

On differentiation of (1.60) with respect to Ψ and noting (1.57), it follows that the M-step requires $\Psi^{(k+1)}$ to be chosen by solving the equation

$$E_{\Psi}\{t(\mathbf{X})\} = t^{(k)}. \quad (1.61)$$

If equation (1.61) can be solved for $\Psi^{(k+1)}$ in Ω , then the solution is unique due to the well-known convexity property of minus the log likelihood of the regular exponential family. In cases where the equation is not solvable, the maximizer $\Psi^{(k+1)}$ of $L(\Psi)$ lies on the boundary of Ω .

1.5.4 Example 1.4: Censored Exponentially Distributed Survival Times (Example 1.3 Continued)

We return now to Example 1.3. It can be seen in this example that the complete-data distribution has the exponential family form (1.56) with natural parameter μ^{-1} and sufficient statistic

$$t(\mathbf{X}) = \sum_{j=1}^n W_j.$$

Hence the E-step requires the calculation of

$$\begin{aligned} t^{(k)} &= E_{\Psi^{(k)}} \{t(\mathbf{X}) \mid \mathbf{y}\} \\ &= \sum_{j=1}^r c_j + \sum_{j=r+1}^n (c_j + \mu^{(k)}) \\ &= \sum_{j=1}^n c_j + (n - r)\mu^{(k)} \end{aligned}$$

from (1.50).

The M-step then yields $\mu^{(k+1)}$ as the value of μ that satisfies the equation

$$\begin{aligned} t^{(k)} &= E_\mu \{t(\mathbf{X})\} \\ &= n\mu. \end{aligned}$$

This latter equation can be seen to be equivalent to (1.52), as derived by direct differentiation of the Q -function $Q(\mu; \mu^{(k)})$.

1.5.5 Generalized EM Algorithm

Often in practice, the solution to the M-step exists in closed form. In those instances where it does not, it may not be feasible to attempt to find the value of Ψ that globally maximizes the function $Q(\Psi; \Psi^{(k)})$. For such situations, DLR defined a generalized EM algorithm (GEM algorithm) for which the M-step requires $\Psi^{(k+1)}$ to be chosen such that

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) \geq Q(\Psi^{(k)}; \Psi^{(k)}) \quad (1.62)$$

holds. That is, one chooses $\Psi^{(k+1)}$ to increase the Q -function $Q(\Psi; \Psi^{(k)})$ over its value at $\Psi = \Psi^{(k)}$, rather than to maximize it over all $\Psi \in \Omega$. As to be shown in Section 3.3, the above condition on $\Psi^{(k+1)}$ is sufficient to ensure that

$$L(\Psi^{(k+1)}) \geq L(\Psi^{(k)}).$$

Hence the likelihood $L(\Psi)$ is not decreased after a GEM iteration, and so a GEM sequence of likelihood values must converge if bounded above. In Section 3.3, we shall discuss what specifications are needed on the process of increasing the Q -function in order to ensure that the limit of $\{L(\Psi^{(k)})\}$ is a stationary value and that the sequence of GEM iterates $\{\Psi^{(k)}\}$ converges to a stationary point.

1.5.6 GEM Algorithm Based on One Newton-Raphson Step

In those situations where the global maximizer of the Q -function $Q(\Psi; \Psi^{(k)})$ does not exist in closed form, consideration may be given to using the Newton-Raphson procedure to iteratively compute $\Psi^{(k+1)}$ on the M-step. As remarked above, it is not essential that $\Psi^{(k+1)}$ actually maximizes the Q -function for the likelihood to be increased. We can use a GEM algorithm where $\Psi^{(k+1)}$ need satisfy only (1.62), which is a sufficient condition to guarantee the monotonicity of the sequence of likelihood values $\{L(\Psi^{(k)})\}$. In some instances, the limiting value $\Psi^{(k+1)}$ of the Newton-Raphson method may not be a global maximizer. But if condition (1.62) is confirmed to hold on each M-step, then at least the user knows that $\{\Psi^{(k)}\}$ is a GEM sequence.

Following Wu (1983) and Jørgensen (1984), Rai and Matthews (1993) propose taking $\Psi^{(k+1)}$ to be of the form

$$\Psi^{(k+1)} = \Psi^{(k)} + a^{(k)} \delta^{(k)}, \quad (1.63)$$

where

$$\delta^{(k)} = -[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}}^{-1} [\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi^{(k)}} \quad (1.64)$$

and where $0 < a^{(k)} \leq 1$.

It can be seen that in the case of $a^{(k)} = 1$, (1.63) is the first iterate obtained when using the Newton-Raphson procedure to obtain a root of the equation

$$\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi = \mathbf{0}.$$

The idea is to choose $a^{(k)}$ so that (1.64) defines a GEM sequence; that is, so that (1.62) holds. It can be shown that

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) - Q(\Psi^{(k)}; \Psi^{(k)}) = a^{(k)} \mathbf{S}(\mathbf{y}; \Psi^{(k)})^T \mathbf{A}^{(k)} \mathbf{S}(\mathbf{y}; \Psi^{(k)}), \quad (1.65)$$

where

$$\mathbf{A}^{(k)} = \mathcal{I}_c^{-1}(\Psi^{(k)}; \mathbf{y}) \{ \mathbf{I}_d - \frac{1}{2} a^{(k)} \tilde{\mathcal{I}}_c^{(k)}(\mathbf{y}) \mathcal{I}_c^{-1}(\Psi^{(k)}; \mathbf{y}) \} \quad (1.66)$$

and where

$$\begin{aligned} \tilde{\mathcal{I}}_c^{(k)}(\mathbf{y}) &= -[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\tilde{\Psi}^{(k)}} \\ &= E_{\Psi^{(k)}} \{ \mathcal{I}_c(\tilde{\Psi}^{(k)}; \mathbf{X}) \mid \mathbf{y} \}, \end{aligned}$$

and $\tilde{\Psi}^{(k)}$ is a point on the line segment from $\Psi^{(k)}$ to $\Psi^{(k+1)}$; \mathbf{I}_d denotes the $d \times d$ identity matrix. Thus the left-hand side of (1.65) is nonnegative if the matrix $\mathbf{A}^{(k)}$ is positive definite.

Typically in practice, $\mathcal{I}_c(\Psi^{(k)}; \mathbf{y})$ is positive definite and so then we have a GEM sequence if the matrix

$$\mathbf{I}_d - \frac{1}{2} a^{(k)} \tilde{\mathcal{I}}_c^{(k)}(\mathbf{y}) \mathcal{I}_c^{-1}(\Psi^{(k)}; \mathbf{y}) \quad (1.67)$$

is positive definite, which can be achieved by choosing the constant $a^{(k)}$ sufficiently small. Suppose that the sequence $\{\Psi^{(k)}\}$ tends to some limit point as $k \rightarrow \infty$. Then it can be seen from (1.66) that, as $k \rightarrow \infty$, $a^{(k)} < 2$ will ensure that (1.67) holds.

The derivation of (1.65) is to be given in Section 4.12, where the use of this GEM algorithm in an attempt to reduce the computation on the M-step, is to be considered further.

1.5.7 EM Gradient Algorithm

The algorithm that uses one Newton-Raphson step to approximate the M-step of the EM algorithm (that is, uses (1.63) with $a^{(k)} = 1$) is referred to by Lange (1995a) as the EM gradient algorithm. It forms the basis of the quasi-Newton approach of Lange (1995b) to speed up the convergence of the EM algorithm, as to be considered in Section 4.14. But as pointed out by Lange (1995b), it is an interesting algorithm in its own right, and is to be considered further in Section 4.13.

1.5.8 EM Mapping

Any instance of the EM (GEM) algorithm as described above implicitly defines a mapping $\Psi \rightarrow M(\Psi)$, from the parameter space of Ψ , Ω , to itself such that

$$\Psi^{(k+1)} = M(\Psi^{(k)}) \quad (k = 0, 1, 2, \dots). \quad (1.68)$$

If $\Psi^{(k)}$ converges to some point Ψ^* and $M(\Psi)$ is continuous, then Ψ^* must satisfy

$$\Psi^* = M(\Psi^*).$$

Thus Ψ^* is a fixed point of the map M .

It is easy to show that if the MLE $\hat{\Psi}$ of Ψ is the unique global maximizer of the likelihood function, then it is a fixed point of the EM algorithm (although there is no guarantee that it is the only one). To see this, we note that the M-step of the EM algorithm (or a GEM algorithm) implies that

$$L(M(\hat{\Psi})) \geq L(\hat{\Psi}). \quad (1.69)$$

Thus $M(\hat{\Psi}) = \hat{\Psi}$, as otherwise (1.69) would contradict the assertion that

$$L(\hat{\Psi}) > L(\Psi)$$

for all Ψ (not equal to $\hat{\Psi}$) $\in \Omega$.

1.6 EM ALGORITHM FOR MAXIMUM A POSTERIORI AND MAXIMUM PENALIZED ESTIMATION

1.6.1 Maximum *a Posteriori* Estimation

The EM algorithm is easily modified to produce the maximum *a posteriori* (MAP) estimate or the maximum penalized likelihood estimate (MPLE) in incomplete-data problems. We consider first the computation of the MAP estimate in a Bayesian framework via the EM algorithm, corresponding to some prior density $p(\Psi)$ for Ψ . We let the incomplete- and complete-data posterior densities for Ψ be given by $p(\Psi | \mathbf{y})$ and $p(\Psi | \mathbf{x})$, respectively. Then the MAP estimate of Ψ is the value of Ψ that maximizes the log (incomplete-data) posterior density which, on ignoring an additive term not involving Ψ , is given by

$$\log p(\Psi | \mathbf{y}) = \log L(\Psi) + \log p(\Psi). \quad (1.70)$$

Here $p(\cdot)$ is being used as a generic symbol for a p.d.f.

The EM algorithm is implemented as follows to compute the MAP estimate.

E-Step. On the $(k+1)$ th iteration, calculate the conditional expectation of the log complete-data posterior density given the observed data vector \mathbf{y} , using the current MAP estimate $\Psi^{(k)}$ of Ψ . That is, calculate

$$E_{\Psi^{(k)}} \{ \log p(\Psi | \mathbf{x}) | \mathbf{y} \} = Q(\Psi; \Psi^{(k)}) + \log p(\Psi). \quad (1.71)$$

M-Step. Choose $\Psi^{(k+1)}$ to maximize (1.71) over $\Psi \in \Omega$.

It can be seen that the E-step is effectively the same as for the computation of the MLE of Ψ in a frequentist framework, requiring the calculation of the Q -function, $Q(\Psi; \Psi^{(k)})$. The M-step differs in that the objective function for the maximization process is equal to $Q(\Psi; \Psi^{(k)})$ augmented by the log prior density, $\log p(\Psi)$. The presence of this latter term as the result of the imposition of a Bayesian prior for Ψ almost always makes the objective function more concave.

1.6.2 Example 1.5: A Multinomial Example (*Example 1.1 Continued*)

We now discuss a Bayesian version of Example 1.1. As before with this example in considering the conditional distribution of the complete-data vector x given the observed data vector y , we can effectively work with the conditional distribution of the missing data vector Z given y . We choose the prior distribution of Ψ to be the beta (ν_1, ν_2) distribution with density

$$p(\Psi) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \Psi^{\nu_1-1} (1 - \Psi)^{\nu_2-1}, \quad (1.72)$$

which is a natural conjugate of the conditional predictive distribution of the missing data. The latter is binomial with sample size y_1 and probability parameter

$$\frac{\frac{1}{4}\Psi}{\frac{1}{2} + \frac{1}{4}\Psi}.$$

From (1.22) and (1.72), we have that

$$\begin{aligned} \log p(\Psi | x) &= \log L(\Psi) + \log p(\Psi) \\ &= (y_{12} + y_4 + \nu_1 - 1) \log \Psi + (y_2 + y_3 + \nu_2 - 1) \log(1 - \Psi), \end{aligned} \quad (1.73)$$

apart from an additive constant. It can be seen from (1.73) that $p(\Psi | x)$ has the beta form. The E-step is effected the same as in the computation of the MLE of Ψ in Example 1.1, with y_{12} in (1.22) replaced by its current conditional expectation

$$y_1 \frac{\frac{1}{4}\Psi^{(k)}}{\frac{1}{2} + \frac{1}{4}\Psi^{(k)}}.$$

On the M-step, the $(k+1)$ th iterate for the MAP estimate is given by

$$\Psi^{(k+1)} = \frac{y_{12}^{(k)} + y_4 + \nu_1 - 1}{y_{12}^{(k)} + y_4 + y_2 + y_3 + \nu_1 + \nu_2 - 2}.$$

Note that this beta distribution is uniform on $[0,1]$ when $\nu_1 = \nu_2 = 1$. When the prior is uniform, the MAP estimate is the same as the MLE of Ψ in the frequentist framework. We return to related techniques for Bayesian estimation in Chapter 6.

1.6.3 Maximum Penalized Estimation

In the case of MPL estimation, $\log p(\Psi)$ in (1.70) is taken to have the form

$$\log p(\Psi) = -\xi K(\Psi), \quad (1.74)$$

where $K(\Psi)$ is a roughness penalty and ξ is a smoothing parameter. Often $K(\Psi)$ is of the form

$$K(\Psi) = \Psi^T \mathbf{A} \Psi.$$

For instance in ridge regression, $\mathbf{A} = \mathbf{I}_d$.

The EM algorithm can be applied in the same manner as above to compute the MPLE of Ψ , where now ξ is an additional parameter to be estimated along with Ψ . In Section 5.17, we consider a modified version of the EM algorithm, the one-step-late (OSL) algorithm as proposed by Green (1990b), that facilitates the computation of the MPLE.

1.7 BRIEF SUMMARY OF THE PROPERTIES OF THE EM ALGORITHM

In the following chapters, we shall discuss in detail various properties of the EM algorithm and some computational aspects of it, and we shall present a number of illustrations and applications of the algorithm. However, in order to give the reader a quick idea of the algorithm's potential as a useful tool in statistical estimation problems, we summarize here the reasons for its appeal. We also mention some of the criticisms leveled against the algorithm.

The EM algorithm has several appealing properties relative to other iterative algorithms such as Newton-Raphson and Fisher's scoring method for finding MLE's. Some of its advantages compared to its competitors are as follows:

1. The EM algorithm is numerically stable, with each EM iteration increasing the likelihood (except at a fixed point of the algorithm).
2. Under fairly general conditions, the EM algorithm has reliable global convergence. That is, starting from an arbitrary point $\Psi^{(0)}$ in the parameter space, convergence is nearly always to a local maximizer, barring very bad luck in the choice of $\Psi^{(0)}$ or some local pathology in the log likelihood function.
3. The EM algorithm is typically easily implemented, because it relies on complete-data computations: the E-step of each iteration only involves taking expectations over complete-data conditional distributions and the M-step of each iteration only requires complete-data ML estimation, which is often in simple closed form.
4. The EM algorithm is generally easy to program, since no evaluation of the likelihood nor its derivatives is involved.
5. The EM algorithm requires small storage space and can generally be carried out on a small computer. For instance, it does not have to store the information matrix nor its inverse at any iteration.
6. Since the complete-data problem is likely to be a standard one, the M-step can often be carried out using standard statistical packages in situations where the complete-data MLE's do not exist in closed form. In other such situations, extensions of the EM algorithm such as the GEM and the expectation-conditional maximization (ECM) algorithms often enable the M-step to be implemented iteratively in a fairly simple manner. Moreover, these extensions share the stable monotone convergence of the EM algorithm.
7. The analytical work required is much simpler than with other methods since only the conditional expectation of the log likelihood for the complete-data problem needs to

be maximized. Although a certain amount of analytical work may be needed to carry out the E-step, it is not complicated in many applications.

8. The cost per iteration is generally low, which can offset the larger number of iterations needed for the EM algorithm compared to other competing procedures.
9. By watching the monotone increase in likelihood (if evaluated easily) over iterations, it is easy to monitor convergence and programming errors.
10. The EM algorithm can be used to provide estimated values of the “missing” data.

Some of the criticisms of the EM algorithm are as follows:

1. Unlike the Fisher’s scoring method, it does not have an inbuilt procedure for producing an estimate of the covariance matrix of the parameter estimates. However, as to be discussed in Section 1.8 and to be pursued further in Chapter 4, this disadvantage can easily be removed by using appropriate methodology associated with the EM algorithm.
2. The EM algorithm may converge slowly even in some seemingly innocuous problems and in problems where there is too much ‘incomplete information’.
3. The EM algorithm like the Newton-type methods does not guarantee convergence to the global maximum when there are multiple maxima. Further, in this case, the estimate obtained depends upon the initial value. But, in general, no optimization algorithm is guaranteed to converge to a global or local maximum, and the EM algorithm is not magical in this regard. There are other procedures such as simulated annealing to tackle such situations. These are, however, complicated to apply.
4. In some problems, the E-step may be analytically intractable, although in such situations there is the possibility of effecting it via a Monte Carlo approach, as to be discussed in Section 6.3.

1.8 HISTORY OF THE EM ALGORITHM

1.8.1 Early EM History

The earliest reference to literature on an EM-type of algorithm is Newcomb (1886), who considers estimation of parameters of a mixture of two univariate normals. McKendrick (1926) gives a medical application of a method in the spirit of the EM algorithm. Meng and van Dyk (1997) in an interesting article capture the essence and spirit of this bit of EM algorithm’s ancient history.

1.8.2 Work Before Dempster, Laird, and Rubin (1977)

The statistical literature is strewn with methods in the spirit of the EM algorithm or which are actually EM algorithms in special contexts. The formulation of the EM algorithm in its present generality is due to DLR, who also give a variety of examples of its applicability and establish its convergence and other basic properties under fairly general conditions. They identify the common thread in a number of algorithms, formulate it in a general setting, and show how some of these algorithms for special problems are special cases of their EM

algorithm. They also point out new applications of the algorithm. In this subsection, we give an account of a few algorithms for special contexts, which have preceded the general formulation of DLR.

Healy and Westmacott (1956) propose an iterative method for estimating a missing value in a randomized block design, which turns out to be an example of the EM algorithm, as to be pointed out in Examples 2.2 and 2.3 in Section 2.3.

Hartley (1958) gives a treatment of the general case of count data and enunciates the basic ideas of the EM algorithm. Buck (1960) considers the estimation of the mean vector and the covariance matrix of a p -dimensional population when some observations on all the p variables together with some observations on some of the variables only are available. He suggests a method of imputation of the missing values by regressing a missing variable on the observed variables in each case, using only cases for which observations on all variables are available. Then the parameters are estimated from the observations ‘completed’ in this way together with a ‘correction’ for the covariance matrix elements. The interesting aspect of Buck’s (1960) method is that all the required regressions and correction terms can be computed by a single operation of inverting the information matrix based on the complete observations and suitable pivoting and sweeping operations. Buck’s (1960) procedure gives the MLE under certain conditions. It also has the basic elements of the EM algorithm. Blight (1970) considers the problem of finding the MLE’s of the parameters of an exponential family from Type I censored sample and derives the likelihood equation. By suitably interpreting the likelihood equation, he derives an iterative method for its solution, which turns out to be an EM algorithm. He also obtains some convergence results and derives the asymptotic covariance matrix of the estimator.

In a series of papers, Baum and Petrie (1966), Baum and Eagon (1967), and Baum, Petrie, Soules, and Weiss (1970) deal with an application of the EM algorithm in a Markov model; this paper contains some convergence results, which generalize easily. Further, the algorithm developed here is the basis of present-day EM algorithms used in hidden Markov models. Orchard and Woodbury (1972) introduce the Missing Information principle, which is very much related to the spirit and basic ideas of the EM algorithm and note the general applicability of the principle. The relationship between the complete- and incomplete-data log likelihood functions established by them, leads to the fact that the MLE is a fixed point of a certain transform. This fact is also noted in many different contexts by a number of authors, who exploit it to develop quite a few forerunners to the EM algorithm. Carter and Myers (1973) consider a special type of mixture of discrete distributions, for which in the case of partially classified data, no closed form solution exists for the MLE. They work out the likelihood equation and derive an algorithm to solve it using Hartley’s (1958) method; this algorithm turns out to be an EM algorithm. Astronomers involved in quantitative work have been familiar with what is known as the Richardson-Lucy algorithm for deconvolution of images for restoring degraded images, based on the work of Richardson (1972) and Lucy (1974); this is an instance of the EM algorithm. Chen and Fienberg (1974) consider the problem of a two-way contingency table with some units classified both ways and some classified by only one of the ways and derive an algorithm for computation of the MLE’s of the cell probabilities, which turns out to be an EM algorithm. Haberman (1974) also deals with the application of an EM-type algorithm in contingency tables with partially classified data. Efron (1967) introduces the so-called Self-Consistency principle in a non-parametric setting for a wide class of incomplete-data problems as an intuitive analog of the ML principle and introduces the Self-Consistency algorithm for right-censored problems. Turnbull (1974) deals with nonparametric estimation of a survivorship function from doubly censored data based on the idea of self-consistency due to Efron (1967) and derives an

iterative procedure. Turnbull (1976) deals with the empirical distribution with arbitrarily grouped, censored, and truncated data, derives a version of the EM algorithm, and notes that not only actually missing data problems, but also problems such as with truncated data, can be treated as incomplete-data problems; he calls individuals who are never observed “ghosts.” Thus Turnbull (1974) extends Efron’s (1967) idea and shows its equivalence with the nonparametric likelihood equations for these problems. He also proves convergence of EM-like algorithms for these problems. Prior to the appearance of the DLR paper, the resolution of mixtures of distributions for a variety of situations and distributional families gave rise to a number of algorithms that can be regarded as particular applications of the EM algorithm. These papers, which are surveyed in McLachlan (1982) and McLachlan and Basford (1988), include the seminal paper of Day (1969). Also, McLachlan (1975, 1977) proposed an iterative method for forming the (normal theory-based) linear discriminant function from partially classified training data.

The basic idea of the EM algorithm is also in use in the “gene-counting” method used by geneticists in the estimation of ABO blood group gene frequencies and other genetic problems (Ceppellini, Siniscalco, and Smith, 1955; Smith, 1957), as noted in Example 2.4 of Section 2.4.

Sundberg (1974, 1976) deals with properties of the likelihood equation in the general context of incomplete-data problems from exponential families, and arrives at special forms for the likelihood equation and the information matrix, which have come to be known as Sundberg formulas. Sundberg (1976) acknowledges that his key “iteration mapping,” which corresponds to the EM mapping of DLR, was suggested to him by A. Martin-Löf in a personal communication in 1967. Beale and Little (1975) develop an algorithm and the associated theory for the multivariate normal case with incomplete data.

All this work was done before DLR formulated the problem in its generality. Indeed, there are many other algorithms found in the literature before DLR which are in the spirit of the EM algorithm or are actually EM algorithms in special contexts. We have not mentioned them all here.

1.8.3 EM Examples and Applications Since Dempster, Laird, and Rubin (1977)

After the general formulation of the EM algorithm by DLR, some well-known ML estimation methods in various contexts have been shown to be EM algorithms, in DLR itself and by others. Iteratively Reweighted Least Squares (IRLS) is an iterative procedure for estimating regression coefficients, wherein each iteration is a weighted least-squares procedure with the weights changing with the iterations. It is applied to robust regression, where the estimates obtained are MLE’s under suitable distributional assumptions. Dempster, Laird, and Rubin (1980) show that the IRLS procedure is an EM algorithm under distributional assumptions. Again, the well-known and standard method of estimating variance components in a mixed model, Henderson’s algorithm, is shown to be an EM-type algorithm by Laird (1982). Gill (1989) observes that in such missing value problems the score functions of suitably chosen parametric submodels coincide exactly with the self-consistency equations and also have an interpretation in terms of the EM algorithm. Wolynetz (1979a, 1979b, 1980) deals with the case of confined and censored data and derives the ML regression line under the assumption that residuals of the dependent variables are normally distributed, using the EM algorithm. It turns out that this line is the same as the “iterative least-squares” line derived by Schmee and Hahn (1979) in an industrial context and the “detections and

bounds” regression developed by Avni and Tananbaum (1986) in the context of a problem in astronomy.

Schlossmacher (1973) proposes an IRLS procedure for computing Least Absolute Deviation (LAD) regression estimates. Pettitt (1985) considers the least-squares estimation problem in linear regression with error distribution as Student’s t and applies the EM algorithm. Student’s t distribution is an example of a scale mixture of a normal distribution. A random variable U with such a mixture has a density of the form

$$\kappa^{-1} \int_0^\infty q\phi\left(\frac{qu}{\kappa}\right)dG(q) \quad (1.75)$$

where $\kappa > 0$, ϕ is the standard normal density, and G is a distribution function (see Andrews, and Mallows, 1974). Another useful example of a scale mixture is the double exponential (Laplace) distribution, which is used as a long-tailed alternative to the normal distribution for modeling error terms. The MLE of regression parameters with double exponentially distributed error terms turns out to be the LAD estimator. Thus it is often used in robust estimation contexts like the t distribution. Phillips (2002) derives the EM algorithm for MLE of linear regression parameters under double exponentially distributed errors; he shows that this is the same as Schlossmacher’s IRLS algorithm for computing LAD estimates. Phillips proposes a slightly modified version of Schlossmacher’s algorithm and a GEM algorithm for LAD estimates of linear and nonlinear regression parameters.

The EM algorithm has been applied to neural networks with hidden units to derive training algorithms; in the M-step, this involves a version of the iterative proportional fitting algorithm for multiway contingency tables (Byrne, 1992; Cheng and Titterington, 1994). Csiszár and Tusnády (1984), Amari, Kurata, and Nagaoka (1992), Byrne (1992), and Amari (1995a, 1995b) explore the connection between the EM algorithm and information geometry. It is pointed out that the EM algorithm is useful in the learning of hidden units in a Boltzmann machine and that the steps of the EM algorithm correspond to the e -geodesic and m -geodesic projections in a manifold of probability distributions, in the sense of statistical inference and differential geometry. The EM algorithm is also useful in the estimation of parameters in hidden Markov models, which are applicable in speech recognition (Rabiner, 1989) and image processing applications (Besag, 1986); these models can be viewed as more general versions of the classical mixture resolution problems for which the EM algorithm has already become a standard tool (Titterington, 1990; Qian and Titterington, 1991).

The DLR paper proved to be a timely catalyst for further research into the applications of finite mixture models. This is witnessed by the subsequent stream of papers on finite mixtures in the literature, commencing with, for example, Ganesalingam and McLachlan (1978, 1979). As Aitkin and Aitkin (1994) note, almost all the post-1978 applications of mixture modeling reported in the books on mixtures by Titterington, Smith, and Makov (1985) and McLachlan and Basford (1988), use the EM algorithm.

1.8.4 Two Interpretations of EM

In the innumerable independent derivations of the EM algorithm for special problems, especially the various versions of the mixture resolution problem, two interpretations are discernible. They are:

1. The EM algorithm arises naturally from the particular forms taken by the derivatives of the log likelihood function. Various authors have arrived at the EM algorithm in special cases, while attempting to manipulate the likelihood equations to be able to solve them in an elegant manner.

2. Many a problem for which the MLE is complex, can be viewed as an incomplete-data problem with a corresponding complete-data problem, suitably formulated, so that the log likelihood functions of the two problems have a nice connection, which can be exploited to arrive at the EM algorithm.

The first interpretation is reflected in the following studies, all of which are on mixture resolution and which preceded DLR. Finite mixtures of univariate normal distributions are treated by Hasselblad (1966) and Behboodian (1970), arbitrary finite mixtures by Hasselblad (1969), mixtures of two multivariate normal distributions with a common covariance matrix by Day (1969), and mixtures of multivariate normal distributions with arbitrary covariance matrices by Wolfe (1967, 1970). This interpretation is also reflected in Blight (1970), who considers exponential families under Type I censoring, in Tan and Chang (1972), who consider a mixture problem in genetics, in Hosmer (1973a) who carries out Monte Carlo studies on small sample sizes with mixtures of two normal distributions, in Hosmer (1973b) who extends his earlier results to the case of a partially classified sample, in the book by Duda and Hart (1973), where the use of multivariate normal mixtures in unsupervised pattern recognition is considered, in Hartley (1978), where a ‘switching regression’ model is considered, and in Peters and Coberly (1976) who consider ML estimation of the proportions in a mixture.

The second interpretation is reflected in the works of Orchard and Woodbury (1972), who were the first to formulate the Missing Information principle and to apply it in various problems, in Healy and Westmacott (1956) and other works on missing values in designed experiments, in Buck (1960) on the estimation of the mean and the covariance matrix of a random vector, in Baum et al. (1970) who consider the general mixture density estimation problem, in Hartley and Hocking (1971) who consider the general problem of analysis of incomplete data, in Haberman (1974, 1976, 1977) who considers log linear models for frequency tables derived by direct and indirect observations, iteratively reweighted least-squares estimation, and product models, and in the works of Cepellini et al. (1955), Chen (1972), Goodman (1974), and Thompson (1975).

1.8.5 Developments in EM Theory, Methodology, and Applications

Dempster, Laird, and Rubin (1977) establish important fundamental properties of the algorithm. In particular, these properties imply that typically in practice the sequence of EM iterates will converge to a local maximizer of the log likelihood function $\log L(\Psi)$. If $L(\Psi)$ is unimodal in Ω with Ψ^* being the only stationary point of $L(\Psi)$, then for any EM sequence $\{\Psi^{(k)}\}$, $\Psi^{(k)}$ converges to the unique maximizer Ψ^* of $L(\Psi)$. In general, if $\log L(\Psi)$ has several (local or global) maxima and stationary values, convergence of the EM sequence to either type depends on the choice of starting point. Furthermore, DLR show that convergence is linear with the rate of convergence proportional to λ_{\max} , where λ_{\max} is the maximal fraction of missing information. This implies that the EM algorithm can be very slow to converge, but then the intermediate values do provide very valuable statistical information. This also implies that the choice of the complete-data problem can influence the rate of convergence, since this choice will determine λ_{\max} .

There have been quite a few developments in the methodology of the EM algorithm since DLR. Wu (1983) gives a detailed account of the convergence properties of the EM algorithm, addressing, in particular, the problem that the convergence of $L(\Psi^{(k)})$ to L^* does not automatically imply the convergence of $\Psi^{(k)}$ to a point Ψ^* . On this same matter, Boyles (1983) presents an example of a generalized EM sequence that converges to the circle

of the unit radius and not to a single point. Horng (1986, 1987) presents many interesting examples of sublinear convergence of the EM algorithm. Lansky, Casella, McCulloch, and Lansky (1992) establish some invariance, convergence, and rates of convergence results. The convergence properties of the EM algorithm are to be pursued further in Section 3.4.

In a series of papers, Turnbull, and Mitchell (1978, 1984) and Mitchell and Turnbull (1979) discuss nonparametric ML estimation in survival/sacrifice experiments and show its self-consistency and convergence. Laird (1978) deals with nonparametric ML estimation of a mixing distribution and points out the equivalence of the Self-Consistency principle and Orchard and Woodbury's (1972) Missing Information principle. Laird (1978) also shows that in the case of parametric exponential families, these two principles have the same mathematical basis as the Sundberg formulas. She also establishes that the self-consistency algorithm is a special case of the EM algorithm.

One of the initial criticisms of the EM algorithm was that unlike Newton-type methods, it does not automatically produce an estimate of the covariance matrix of the MLE. In an important development associated with the EM methodology, Louis (1982) develops a method of finding the observed information matrix while using the EM algorithm, which is generally applicable. This method gives the observed information matrix in terms of the gradient and curvature of the complete-data log likelihood function, which is more amenable to analytical calculations than the incomplete-data analog. Fisher (1925) had observed the result that the incomplete-data score statistic is the conditional expected value of the complete-data score statistic given the incomplete observations (observed data). Efron (1977) in his comments on DLR connects Fisher's result with incompleteness. Louis (1982) makes this connection deeper by establishing it for the second moments. In other related work, Meilijson (1989) proposes a method of numerically computing the covariance matrix of the MLE, using the ingredients computed in the E- and M-steps of the algorithm, as well as a method to speed up convergence. His method of approximation avoids having to calculate second-order derivatives as with Louis' method. Meilijson (1989) shows that the single-observation scores for the incomplete-data model are obtainable as a by-product of the E-step. He also notes that the expected information matrix can be estimated consistently by the empirical covariance matrix of the individual scores. Previously, Redner and Walker (1984) in the context of the mixture resolution problem, suggested using the empirical covariance matrix of the individual scores to estimate consistently the expected information matrix.

The use of the empirical information matrix as discussed above is of course applicable only in the special case of i.i.d. data. For the general case, Meng and Rubin (1991) define a procedure that obtains a numerically stable estimate of the asymptotic covariance matrix of the EM-computed estimate, using only the code for computing the complete-data covariance matrix, the code for the EM algorithm itself, and the code for standard matrix operations. In particular, neither likelihoods, nor partial derivatives of likelihoods nor log likelihoods need to be evaluated. They refer to this extension of the EM algorithm as the Supplemented EM algorithm.

Baker (1992) reviews methods for computing standard errors in the context of EM computations known up to that point of time. Two numerical differentiation approaches can be discerned in the literature: (1) differentiation of Fisher score vector to obtain the Hessian of log likelihood; (2) differentiation of the EM operator and use of an identity relating derivative to Hessian of log likelihood. The SEM method uses the second approach. Baker (1992) notes the possibility of numerical inaccuracy. SEM requires fairly accurate estimates of the parameters and so the SEM estimates can be expensive (Segal, Bacchetti, and Jewell, 1994; McCulloch, 1998). Moreover, SEM estimates can be numerically unstable. Jamshidian and

Jennrich (2000) suggest three methods, including one of type (1) above by a Richardson extrapolation and one of type (2) above by forward difference and Richardson extrapolation methods. Oakes (1999) facilitates standard error computation in the EM context by deriving a formula for the observed information matrix; he derives an explicit formula for the second derivatives matrix of the observed data log likelihood in terms of the derivatives of the conditional expectation function of the complete data log likelihood given data.

Louis (1982) also suggests a method of speeding up convergence of the EM algorithm using the multivariate generalization of the Aitken acceleration procedure. The resulting algorithm is essentially equivalent to using the Newton-Raphson method to find a zero of the (incomplete-data) score statistic. Jamshidian and Jennrich (1993) use a generalized conjugate gradient approach to accelerate convergence of the EM algorithm. However, attempts to speed up the EM algorithm do reduce its simplicity and there is no longer any guarantee that the likelihood will always increase from one iteration to the next. These points are to be taken up in Chapter 4.

As noted earlier, one of the major reasons for the popularity of the EM algorithm is that the M-step involves only complete-data ML estimation, which is often computationally simple. But if the complete-data ML estimation is rather complicated, then the EM algorithm is less attractive because the M-step is computationally unattractive. In many cases, however, complete-data ML estimation is relatively simple if maximization is undertaken conditional on some of the parameters (or some functions of the parameters). To this end, Meng and Rubin (1993) introduce a class of generalized EM algorithms, which they call the expectation-conditional maximization (ECM) algorithm. The ECM algorithm takes advantage of the simplicity of complete-data conditional maximization by replacing a complicated M-step of the EM algorithm with several computationally simpler CM-steps. Each of these CM-steps maximizes the expected complete-data log likelihood function found in the preceding E-step subject to constraints on Ψ , where the collection of all constraints is such that the maximization is over the full parameter space of Ψ . Liu and Rubin (1994) give a generalization of the ECM algorithm that replaces some of the CM-steps with steps that maximize the constrained actual (incomplete-data) log likelihood. They call this algorithm, the expectation-conditional maximization either (ECME) algorithm. It shares with both the EM and ECM algorithms, their stable monotone convergence and basic simplicity of implementation relative to faster converging competitors. In a further extension, Meng and van Dyk (1997) propose generalizing the ECME algorithm and the SAGE algorithm of Fessler and Hero (1994) by combining them into one algorithm, called the Alternating ECM (AECM) algorithm. It allows the specification of the complete data to be different on each CM-step.

Meng and van Dyk (1997) also consider the problem of speeding up convergence of the EM algorithm. Their approach is through the choice of the missing-data in the specification of the complete-data problem in the EM framework. They introduce a working parameter in the specification of the complete data, which thus indexes a class of EM algorithms. The aim is to select a value of the working parameter that increases the speed of convergence without appreciably affecting the stability and simplicity of the resulting EM algorithm.

In other developments, there is the work of Lange (1995a, 1995b) on the use of the EM gradient algorithm in situations where the solution to the M-step does not exist in closed form. As discussed in Section 1.5.7 and to be considered further in Chapter 4, the EM gradient algorithm approximates the M-step by one Newton-Raphson step. Lange (1995b) subsequently uses the EM gradient algorithm to form the basis of a quasi-Newton approach to accelerate convergence of the EM algorithm. In another development, Heyde and Morton

(1996) extend the EM algorithm to deal with estimation via general estimating functions and in particular the quasi-score.

Parameter Expanded EM (PX-EM) proposed by Liu, Rubin, and Wu (1998) is a method for accelerating the EM algorithm by expanding the parameter space over which the maximization is carried out. This space includes parameters the values of which are known. This often results in speeding up convergence. This idea is related to efficient data augmentation in respect of the missing data structure.

Ruud (1991) reviews the applications of EM from an econometric point of view; he also discusses what might be called Monte Carlo EM. Brockwell and Davis (2002) illustrate an application of the EM algorithm to estimating an AR(2) model with missing observations. Other earlier time series and econometric applications can be found in Shumway and Stoffer (1982) and Watson and Engle (1983).

Cowell, Dawid, Lauritzen, and Spiegelhalter (2003) discuss the EM algorithm in the context of probabilistic networks and expert systems in the presence of incomplete data, in terms of directed acyclic graphs. Thiesson (1997) discusses acceleration of the EM in Bayesian networks in the presence of incomplete data. Thiesson (1995) discusses GEM algorithms in recursive graphical association models. Geng, Asano, Ichimura, Tao, Wan, and Kuroda (1996) discuss partial imputation in the EM algorithm. Didelez and Pigeot (1998) discuss MLE in graphical models with missing values.

One of the directions in which EM-related algorithms have been extended is Monte Carlo. Stochastic EM, Monte Carlo EM, Imputation methods, and Markov chain Monte Carlo are some examples of these algorithms. We present a whole chapter on these Monte Carlo versions of the EM algorithm, wherein we include a historical review of these algorithms.

The key idea of the EM algorithm where a surrogate function of the log likelihood is maximized in a iterative procedure occurs in quite a few other optimization procedures as well, leading to a more general way of looking at EM as an optimization procedure. We discuss these procedures, along with a historical account of them in another chapter on Generalizations of the EM Algorithm.

1.9 OVERVIEW OF THE BOOK

In Chapter 2, the EM methodology presented in this chapter is illustrated in some commonly occurring situations such as missing observations in multivariate normal data sets and regression problems, the multinomial distribution with complex cell-probability structure, grouped data, data from truncated distributions, and the fitting of finite mixture models.

The basic theory of the EM algorithm is presented in Chapter 3. In particular, the convergence properties and the rates of convergence are systematically examined. Consideration is given also to the associated Missing-Information principle.

In Chapter 4, two important issues associated with the use of the EM algorithm are considered, namely the provision of standard errors and the speeding up of its convergence. In so doing, we discuss several of the many modifications, extensions, and alternatives to the EM methodology that appear in the literature.

We discuss further modifications and extensions to the EM algorithm in Chapter 5. In particular, the extensions of the EM algorithm known as the smoothed EM, ECM, multi-cycle ECM, ECME, and AECM algorithms are given. We also present the EM gradient algorithm and a consequent quasi-Newton approach to accelerate its convergence. Having presented the easier illustrations in Chapter 2, the more difficult problems that motivated the development of these extensions are illustrated in Chapter 5, including estimation for

variance components, linear mixed models, repeated-measures designs, factor analysis, and principal component analysis.

In Chapter 6, we explore various Monte Carlo variations and versions of the EM algorithm. In the process, we present a concise account of the standard independent and identically distributed (i.i.d.) Monte Carlo algorithms like Rejection Sampling and its variations, the techniques of Monte Carlo integration, and Markov chain Monte Carlo (MCMC) algorithms of Metropolis-Hastings and Gibbs Sampling. We discuss Monte Carlo EM, Stochastic EM, Bayesian EM, and other such extensions of EM, some of which are useful in the context of intractability of the E-step and others in the Bayesian context of computing the Maximum a Posteriori (MAP) estimate. We discuss Data Augmentation and Multiple Imputation. We then establish several connections between the EM algorithm in the frequentist context to MCMC in Bayesian contexts. We present many examples.

In Chapter 7, we present a few generalizations of the EM algorithm, like an EM algorithm for estimating equations, Variational EM algorithm, and optimization algorithms like the MM algorithm which like the EM find a surrogate function to optimize in an iterative scheme.

The concluding chapter, Chapter 8, discusses a few applications like in Hidden Markov Models, Neural Networks, and AIDS epidemiology.

1.10 NOTATIONS

We now define the notations that are used consistently throughout the book. Less frequently used notations will be defined later when they are first introduced.

All vectors and matrices are in boldface. The superscript T denotes the transpose of a vector or matrix. The trace of a matrix \mathbf{A} is denoted by $tr(\mathbf{A})$, while the determinant of \mathbf{A} is denoted by $|A|$. The null vector is denoted by $\mathbf{0}$. The notation $\text{diag}(a_1, \dots, a_n)$ is used for a matrix with diagonal elements a_1, \dots, a_n and all off-diagonal elements zero.

Generally, the vector \mathbf{x} is used to represent the so-called complete data, while the vector \mathbf{y} represents the actual observed data (incomplete data). However, in contexts not relating to incomplete data or complete data, \mathbf{x} and \mathbf{y} may not be used in this sense. Where possible, a random vector is represented by the corresponding upper case of the letter used for a particular realization. In this instance, \mathbf{X} and \mathbf{Y} denote the complete- and incomplete-data random vectors corresponding to \mathbf{x} and \mathbf{y} , respectively.

The incomplete-data random vector \mathbf{Y} is taken to be of p -dimensions, having probability density function (p.d.f.) $g(\mathbf{y}; \Psi)$ on \mathbb{R}^p , where

$$\Psi = (\Psi_1, \dots, \Psi_d)^T$$

is the vector containing the unknown parameters in the postulated form for the p.d.f. of \mathbf{Y} . The parameter space is denoted by Ω . In the case where \mathbf{Y} is discrete, we can still view $g(\mathbf{y}; \Psi)$ as a density by the adoption of counting measure.

The likelihood function for Ψ formed from the observed data \mathbf{y} is denoted by

$$L(\Psi) = g(\mathbf{y}; \Psi),$$

while $\log L(\Psi)$ denotes the log likelihood function.

The p.d.f. of the complete-data vector \mathbf{X} is denoted by $g_c(\mathbf{x}; \Psi)$, with

$$L_c(\Psi) = g_c(\mathbf{x}; \Psi)$$

denoting the complete-data likelihood function for Ψ that could be formed from \mathbf{x} if it were completely observable.

The (incomplete-data) score statistic is given by

$$S(\mathbf{y}; \Psi) = \partial \log L(\Psi) / \partial \Psi,$$

while

$$S_c(\mathbf{x}; \Psi) = \partial \log L_c(\Psi) / \partial \Psi$$

denotes the corresponding complete-data score statistic.

The conditional p.d.f. of \mathbf{X} given \mathbf{y} is denoted by

$$k(\mathbf{x} | \mathbf{y}; \Psi).$$

The so-called missing data is represented by the vector \mathbf{z} .

The sequence of EM iterates is denoted by $\{\Psi^{(k)}\}$, where $\Psi^{(0)}$ denotes the starting value of Ψ , and $\Psi^{(k)}$ denotes the value of Ψ on the k th subsequent iteration of the EM algorithm. Such superfixes are also used for components of Ψ and other parameters derived from them; for instance, if μ_1 is a component of Ψ and $\sigma_{22.1}^{(k)}$ is a parameter defined as a function of the components of Ψ , the notations $\mu_1^{(k)}$ and $\sigma_{22.1}^{(k)}$ respectively denote their k th EM iterates.

The Q -function is used to denote the conditional expectation of the complete-data log likelihood function, $\log L_c(\Psi)$, given the observed data \mathbf{y} , using the current fit for Ψ . Hence on the $(k+1)$ th iteration of the E -step, it is given by

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log L_c(\Psi) | \mathbf{y} \},$$

where the expectation operator E has the subscript $\Psi^{(k)}$ to explicitly convey that this (conditional) expectation is being effected using $\Psi^{(k)}$ for Ψ . Concerning other moment operators, we shall use $\text{var}_{\Psi}(W)$ for the variance of a random variable W and $\text{cov}_{\Psi}(W)$ for the covariance matrix of a random vector W , where Ψ is the parameter vector indexing the distribution of W (W).

The maximum likelihood estimate (MLE) of Ψ is denoted by $\hat{\Psi}$.

The (incomplete-data) observed information matrix is denoted by $\mathbf{I}(\hat{\Psi}; \mathbf{y})$, where

$$\mathbf{I}(\Psi; \mathbf{y}) = -\partial^2 \log L(\Psi) / \partial \Psi \partial \Psi^T.$$

The (incomplete-data) expected information matrix is denoted by $\mathcal{I}(\Psi)$, where

$$\mathcal{I}(\Psi) = E_{\Psi} \{ \mathbf{I}(\Psi; \mathbf{Y}) \}.$$

For the complete data, we let

$$\mathbf{I}_c(\Psi; \mathbf{x}) = -\partial^2 \log L_c(\Psi) / \partial \Psi \partial \Psi^T,$$

while its conditional expectation given \mathbf{y} is denoted by

$$\mathcal{I}_c(\Psi; \mathbf{y}) = E_{\Psi} \{ \mathbf{I}_c(\Psi; \mathbf{X}) | \mathbf{y} \}.$$

The expected information matrix corresponding to the complete data is given by

$$\mathcal{I}_c(\Psi) = E_{\Psi} \{ \mathbf{I}_c(\Psi; \mathbf{X}) \}.$$

The so-called missing information matrix is denoted by $\mathcal{I}_m(\Psi; \mathbf{y})$ and is defined as

$$\mathcal{I}_m(\Psi; \mathbf{y}) = -E_{\Psi}\{\partial^2 k(\mathbf{x} | \mathbf{y}; \Psi)/\partial\Psi\partial\Psi^T | \mathbf{y}\}.$$

In other notations involving I , the symbol \mathbf{I}_d is used to denote the $d \times d$ identity matrix, while $I_A(x)$ denotes the indicator function that is 1 if x belongs to the set A and is zero otherwise.

The p.d.f. of a random vector \mathbf{W} having a p -dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ is denoted by $\phi(\mathbf{w}; \boldsymbol{\mu}, \Sigma)$, where

$$\phi(\mathbf{w}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{w} - \boldsymbol{\mu})\}.$$

The notation $\phi(w; \mu, \sigma^2)$ is used to denote the p.d.f. of a univariate normal distribution with mean μ and variance σ^2 , and $\Phi(w; \mu, \sigma^2)$, its cumulative distribution function (c.d.f.).

This Page Intentionally Left Blank

CHAPTER 2

EXAMPLES OF THE EM ALGORITHM

2.1 INTRODUCTION

Before we proceed to present the basic theory underlying the EM algorithm, we give in this chapter a variety of examples to demonstrate how the EM algorithm can be conveniently applied to find the MLE in some commonly occurring situations.

The first three examples concern the application of the EM algorithm to problems where there are missing data in the conventional sense. The other examples relate to problems where an incomplete-data formulation and the EM algorithm derived thereby result in an elegant way of computing MLE's. As to missing-data problems, although they are one of the most important classes of problems where the EM algorithm is profitably used, we do not discuss them further in this book except as continuation of these examples in view of the availability of an excellent book by Little and Rubin (1987,2002), which is devoted to the topic and covers the application of the EM algorithm in that context. Discussions of the missing value problem in a general setting and in the multivariate normal case may also be found in Little (1983a, 1983b, 1993, 1994), and Little and Rubin (1989, 1990).

2.2 MULTIVARIATE DATA WITH MISSING VALUES

2.2.1 Example 2.1: Bivariate Normal Data with Missing Values

Let $\mathbf{W} = (W_1, W_2)^T$ be a bivariate random vector having a normal distribution

$$\mathbf{W} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (2.1)$$

with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

The bivariate normal density to be fitted is given by

$$\phi(\mathbf{w}; \boldsymbol{\Psi}) = (2\pi)^{-1} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\}, \quad (2.2)$$

where the vector of parameters $\boldsymbol{\Psi}$ is given by

$$\boldsymbol{\Psi} = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^T.$$

Suppose we wish to find the MLE of $\boldsymbol{\Psi}$ on the basis of a random sample of size n taken on \mathbf{W} , where the data on the i th variate W_i are missing in m_i of the units ($i = 1, 2$). We label the data so that $\mathbf{w}_j = (w_{1j}, w_{2j})^T$ ($j = 1, \dots, m$) denote the fully observed data points, where $m = n - m_1 - m_2$, w_{2j} ($j = m+1, \dots, m+m_1$) denote the m_1 observations with the values of the first variate w_{1j} missing, and w_{1j} ($j = m+m_1+1, \dots, n$) denote the m_2 observations with the values of the second variate w_{2j} missing.

It is supposed that the “missingness” can be considered to be completely random, so that the observed data can be regarded as a random sample of size m from the bivariate normal distribution (2.1) and an independent pair of independent random samples of size m_i from the univariate normal distributions

$$W_i \sim N(\mu_i, \sigma_{ii}) \quad (2.3)$$

for $i = 1, 2$.

The observed data are therefore given by

$$\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_m^T, \mathbf{v}^T)^T,$$

where the vector \mathbf{v} is given by

$$\mathbf{v} = (w_{2,m+1}, \dots, w_{2,m+m_1}, w_{1,m+m_1+1}, \dots, w_{1,n})^T.$$

The log likelihood function for $\boldsymbol{\Psi}$ based on the observed data \mathbf{y} is

$$\begin{aligned} \log L(\boldsymbol{\Psi}) &= -\{m + \frac{1}{2}(m_1 + m_2)\log(2\pi)\} - \frac{1}{2}m \log |\boldsymbol{\Sigma}| \\ &\quad - \frac{1}{2} \sum_{j=1}^m (\mathbf{w}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{w}_j - \boldsymbol{\mu}) - \frac{1}{2}(m_1 \log \sigma_{22} + m_2 \log \sigma_{11}) \\ &\quad - \frac{1}{2}\{\sigma_{11}^{-1} \sum_{j=m+m_1+1}^n (w_{1j} - \mu_1)^2 + \sigma_{22}^{-1} \sum_{j=m+1}^{m+m_1} (w_{2j} - \mu_2)^2\}. \end{aligned} \quad (2.4)$$

An obvious choice for the complete data here are the n bivariate observations. The complete-data vector \mathbf{x} is then given by

$$\mathbf{x} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T,$$

for which the missing-data vector \mathbf{z} is

$$\mathbf{z} = (w_{1,m+1}, \dots, w_{1,m+m_1}, w_{2,m+m_1+1}, \dots, w_{2,n})^T.$$

The complete-data log likelihood function for Ψ is

$$\begin{aligned} \log L_c(\Psi) &= -n \log(2\pi) - \frac{1}{2}n \log |\Sigma| - \frac{1}{2} \sum_{j=1}^n (\mathbf{w}_j - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{w}_j - \boldsymbol{\mu}) \\ &= -n \log(2\pi) - \frac{1}{2}n \log \xi \\ &\quad - \frac{1}{2}\xi^{-1} [\sigma_{22}T_{11} + \sigma_{11}T_{22} - 2\sigma_{12}T_{12} \\ &\quad - 2\{T_1(\mu_1\sigma_{22} - \mu_2\sigma_{12}) + T_2(\mu_2\sigma_{11} - \mu_1\sigma_{12})\} \\ &\quad + n(\mu_1^2\sigma_{22} + \mu_2^2\sigma_{11} - 2\mu_1\mu_2\sigma_{12})], \end{aligned} \tag{2.5}$$

where

$$T_i = \sum_{j=1}^n w_{ij} \quad (i = 1, 2), \tag{2.6}$$

$$T_{hi} = \sum_{j=1}^n w_{hj}w_{ij} \quad (h, i = 1, 2), \tag{2.7}$$

and where

$$\xi = \sigma_{11}\sigma_{22}(1 - \rho^2)$$

and

$$\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})^{\frac{1}{2}}$$

is the correlation between W_1 and W_2 .

It can be seen that $L_c(\Psi)$ belongs to the regular exponential family with sufficient statistic

$$\mathbf{T} = (T_1, T_2, T_{11}, T_{12}, T_{22})^T.$$

If the complete-data vector \mathbf{x} were available, then the (complete-data) MLE of Ψ , $\hat{\Psi}$, would be easily computed. From the usual results for (complete-data) ML estimation for the bivariate normal distribution, $\hat{\Psi}$ is given by

$$\hat{\mu}_i = T_i/n \quad (i = 1, 2), \tag{2.8}$$

$$\hat{\sigma}_{hi} = (T_{hi} - n^{-1}T_hT_i)/n \quad (h, i = 1, 2). \tag{2.9}$$

We now consider the E-step on the $(k+1)$ th iteration of the EM algorithm, where $\Psi^{(k)}$ denotes the value of Ψ after the k th EM iteration. It can be seen from (2.5) that in order to compute the current conditional expectation of the complete-data log likelihood,

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log L_c(\Psi) | \mathbf{y} \},$$

we require the current conditional expectations of the sufficient statistics T_i and T_{hi} ($h, i = 1, 2$). Thus in effect we require

$$E_{\Psi^{(k)}}(W_{1j} \mid w_{2j})$$

and

$$E_{\Psi^{(k)}}(W_{1j}^2 \mid w_{2j})$$

for $j = m + 1, \dots, m + m_1$, and

$$E_{\Psi^{(k)}}(W_{2j} \mid w_{1j})$$

and

$$E_{\Psi^{(k)}}(W_{2j}^2 \mid w_{1j})$$

for $j = m + m_1 + 1, \dots, n$.

From the well-known properties of the bivariate normal distribution, the conditional distribution of W_2 given $W_1 = w_1$ is normal with mean

$$\mu_2 + \sigma_{12}\sigma_{11}^{-1}(w_1 - \mu_1)$$

and variance

$$\sigma_{22.1} = \sigma_{22}(1 - \rho^2).$$

Thus

$$E_{\Psi^{(k)}}(W_{2j} \mid w_{1j}) = w_{2j}^{(k)}, \quad (2.10)$$

where

$$w_{2j}^{(k)} = \mu_2^{(k)} + (\sigma_{12}^{(k)} / \sigma_{11}^{(k)})(w_{1j} - \mu_1^{(k)}), \quad (2.11)$$

and

$$E_{\Psi^{(k)}}(W_{2j}^2 \mid w_{1j}) = w_{2j}^{(k)2} + \sigma_{22.1}^{(k)} \quad (2.12)$$

for $j = m + m_1 + 1, \dots, n$. Similarly, $E_{\Psi^{(k)}}(W_{1j} \mid w_{2j})$ and $E_{\Psi^{(k)}}(W_{1j}^2 \mid w_{2j})$ are obtained by interchanging the subscripts 1 and 2 in (2.10) to (2.12).

Note that if we were to simply impute the $w_{2j}^{(k)}$ for the missing w_{2j} in the complete-data log likelihood (and likewise for the missing w_{1j}), we would not get the same expression as the Q -function yielded by the E-step, because of the omission of the term $\sigma_{22.1}^{(k)}$ on the right-hand side of (2.12).

The M -step on the $(k+1)$ th iteration is implemented simply by replacing T_i and T_{hi} by $T_i^{(k)}$ and $T_{hi}^{(k)}$, respectively, where the latter are defined by replacing the missing w_{ij} and w_{ij}^2 with their current conditional expectations as specified by (2.10) and (2.12) for $i = 2$ and by their corresponding forms for $i = 1$. Accordingly, $\Psi^{(k+1)}$ is given by

$$\mu_i^{(k+1)} = T_i^{(k)} / n \quad (i = 1, 2), \quad (2.13)$$

$$\sigma_{hi}^{(k+1)} = (T_{hi}^{(k)} - n^{-1}T_h^{(k)}T_i^{(k)}) / n \quad (h, i = 1, 2). \quad (2.14)$$

2.2.2 Numerical Illustration

To illustrate the application of the EM algorithm to this type of problem, we now apply it to the data set below from Rubin (1987). Here, there are $n = 10$ bivariate observations in $m_2 = 2$ of which the value of the second variate w_2 is missing (indicated by ?), but there are no cases of values of the first variate w_1 missing; thus $m_1 = 0$.

Variate 1:	8	11	16	18	6	4	20	25	9	13
Variate 2:	10	14	16	15	20	4	18	22	?	?

In this case, where only one of the variates is incompletely observed, explicit expressions exist for the (incomplete-data) MLE's of the components of Ψ . They are obtained by factoring the (incomplete-data) likelihood function $L(\Psi)$ into one factor corresponding to the marginal density of the variate completely observed (here W_1) and a factor corresponding to the conditional density of W_2 given w_1 ; see, for example, Little and Rubin (1987, 2002, pages 135–136). The explicit formulas for the MLE $\hat{\Psi}$ so obtained are

$$\begin{aligned}\hat{\mu}_1 &= \sum_{j=1}^n w_{1j}/n, \\ \hat{\sigma}_{11} &= \sum_{j=1}^n (w_{1j} - \hat{\mu}_1)^2/n, \\ \hat{\mu}_2 &= \bar{w}_2 + \hat{\beta}(\hat{\mu}_1 - \bar{w}_1), \\ \hat{\sigma}_{22} &= s_{22} + \hat{\beta}^2(\hat{\sigma}_{11} - s_{11}), \\ \hat{\sigma}_{12} &= \hat{\beta}\hat{\sigma}_{11} \\ \hat{\beta} &= s_{12}/s_{11},\end{aligned}$$

where

$$\bar{w}_i = \sum_{j=1}^m w_{ij}/m \quad (i = 1, 2)$$

and

$$s_{hi} = \sum_{j=1}^m (w_{hj} - \bar{y}_h)(w_{ij} - \bar{y}_i)/m \quad (h, i = 1, 2).$$

Results of the EM algorithm so applied to the bivariate data above are given in Table 2.1, where the entries for $k = \infty$ correspond to the MLE's computed from their explicit formulas. Since the first variate is completely observed in this illustration, $T_1^{(k)} = T_1$ and $T_{11}^{(k)} = T_{11}$ for each k , and so $\mu_1^{(k+1)}$ and $\sigma_{11}^{(k+1)}$ remain the same over the EM iterations.

It can be seen that in this simple illustration, the EM algorithm has essentially converged after a few iterations. The starting value $\Psi^{(0)}$ here was obtained by computing the estimates of the parameters on the basis of the m completely recorded data points, so far as W_2 is concerned.

2.2.3 Multivariate Data: Buck's Method

The spirit of the EM algorithm above is captured in a method proposed by Buck (1960) for the estimation of the mean μ and the covariance matrix Σ of a p -dimensional random vector W based on a random sample with observations missing on some variables in some

Table 2.1 Results of the EM Algorithm for Example 2.1 (Missing Data on One Variate).

Iteration	$\mu_1^{(k)}$	$\mu_2^{(k)}$	$\sigma_{11}^{(k)}$	$\sigma_{12}^{(k)}$	$\sigma_{22}^{(k)}$	$-2 \log L(\Psi^{(k)})$
0	13.000	14.87500	40.200	24.93750	28.85937	110.6344
1	13.000	14.62687	40.200	20.94254	26.31958	110.1642
2	13.000	14.61699	40.200	20.88279	26.65277	110.1602
3	13.000	14.64561	40.200	20.888436	26.73346	110.1601
4	13.000	14.61532	40.200	20.88497	26.74991	110.1601
5	13.000	14.61525	40.200	20.88512	26.75322	110.1601
6	13.000	14.61524	40.200	20.88515	26.75389	110.1601
7	13.000	14.61524	40.200	20.88515	26.75402	110.1601
8	13.000	14.61523	40.200	20.88516	26.75405	110.1601
∞	13.000	14.61523	40.200	20.88516	26.75405	110.1601

data points. Buck's method is as follows:

Step 1. Estimate μ and Σ by the sample mean \bar{w} and the sample covariance matrix S using only completely observed cases.

Step 2. Estimate each missing observation by using the linear regression of the variable concerned on the variables on which observations have been made for that case. For this purpose, compute the required linear regressions by using \bar{w} and S . Thus

$$\hat{w}_{ij} = \begin{cases} w_{ij} & \text{if } w_{ij} \text{ is observed,} \\ \bar{w}_i + \sum_l \hat{\beta}_{il}^j (w_{lj} - \bar{w}_l) & \text{if } w_{ij} \text{ is missing,} \end{cases}$$

where the $\hat{\beta}_{il}^j$ are the coefficients in the regression as explained above.

Step 3. Compute the mean vector \bar{w}^* and corrected sum of squares and products matrix A using the 'completed' observations \hat{w}_{ij} .

Step 4. Calculate the final estimates as follows:

$$\hat{\mu} = \bar{w}^*; \quad \hat{\sigma}_{hi} = (A_{hi} + \sum_{j=1}^n c_{hij})/n,$$

where the correction term c_{hij} is the residual covariance of w_h and w_i given the variables observed in the j th data point, if both w_{hj} and w_{ij} are missing and zero otherwise. This can be calculated from S .

The EM algorithm applied under the assumption of multivariate normality for this problem is just an iterative version of Buck's (1960) method.

2.3 LEAST SQUARES WITH MISSING DATA

2.3.1 Healy–Westmacott Procedure

Suppose we are interested in estimating by least squares the parameters of a linear model

$$y_j = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_j + e_j \quad (j = 1, \dots, n) \quad (2.15)$$

when some of the observations y_1, \dots, y_n are missing. In (2.15), \mathbf{x}_j is a p -dimensional vector containing the values of the design variables corresponding to the j th response variable y_j , and e_j denotes the error term which has mean zero and variance σ^2 ($j = 1, \dots, n$).

The model parameters may be regression coefficients or various effects in a designed experiment. In designed experiments, the \mathbf{x}_j -values at which the experiment is conducted are fixed and the response y_j is observed at these points. Here the model assumed is on the conditional distribution of Y_j , the random variable corresponding to y_j , given \mathbf{x}_j for $j = 1, \dots, n$ and they are assumed to be independent; the missing cases do not contribute information on the model parameters unlike in the previous example. However, in designed experiments, the experimental values of \mathbf{x}_j are chosen to make the computations simple. If some y_j -values are missing, then least-squares computations get complicated. An approach of the following sort was suggested by Healy and Westmacott (1956), which exploits the simplicity of the least-squares computations in the complete-data case, in carrying out least-squares computations for the incomplete case:

1. Select trial values for all missing values.
2. Perform the complete-data analysis, that is, compute least-squares estimates of model parameters by the complete-data method.
3. Predict missing values using the estimates obtained above.
4. Substitute predicted values for missing values.
5. Go to Step 2 and continue until convergence of missing values or the residual sum of squares (RSS) is reached.

2.3.2 Example 2.2: Linear Regression with Missing Dependent Values

Let us apply this method to the analysis of a 3^2 experiment with missing values. For $j = 1, \dots, n$, the response variable y_j is the number of lettuce plants at the j th combined level of two factors, nitrogen (x_{1j}) and phosphorus (x_{2j}). The three levels of both nitrogen and phosphorus are denoted by $-1, 0, 1$. Responses corresponding to $(-1, -1)^T$ and $(0, 1)^T$ are missing. A data set adapted from Cochran and Cox (1957) is given in Table 2.2. Suppose we wish to estimate the parameters of the linear regression

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j,$$

where the common variance σ^2 of the error terms e_j is to be estimated by the residual mean square.

From the data of a full 3^2 experiment, least-squares estimates are easily computed as follows:

$$\hat{\beta}_0 = \bar{y}; \quad \hat{\beta}_{(1)} = \frac{y_{+(1)} - y_{-(1)}}{6}; \quad \hat{\beta}_2 = \frac{y_{+(2)} - y_{-(2)}}{6},$$

where \bar{y} is the mean of the nine observations, $y_{+(i)}$ and $y_{-(i)}$ are the total of the responses at which x_{ij} is $+1$ and -1 , respectively, for $i = 1, 2$. The least-squares estimates in the

Table 2.2 Number of Lettuce Plants Emerging in a Partial 3^2 Experiment (Example 2.2).

Nitrogen level (x_{1j})	Phosphorus level (x_{2j})	No. of lettuce plants (y_j)
-1	0	409
-1	1	341
0	-1	413
0	0	358
1	-1	326
1	0	291
1	1	312

Source: Adapted from Cochran and Cox (1957)

incomplete-data case are somewhat more inelegant and complicated. Starting from initial estimates of 358, -51, and -55 for β_0 , β_1 , and β_2 , respectively, computed from the data points $(0, 0)^T$, $(-1, 0)^T$, and $(0, -1)^T$, the Healy–Westmacott procedure yields results given in Table 2.3.

Table 2.3 Results of Healy–Westmacott Procedure on Data of Example 2.2.

Iteration	β_0	β_1	β_2	$\hat{y}(-1, -1)$	$\hat{y}(0, -1)$	RSS
1	357.4445	-47.50000	-41.16667	464.0000	303.0000	1868.158
2	356.9321	-44.51854	-35.97223	446.1111	316.2778	1264.418
3	356.4870	-43.07048	-33.74382	437.4229	320.9599	1045.056
4	356.2272	-42.38354	-32.75969	433.3013	322.7431	955.8019
5	356.0931	-42.06175	-32.31716	431.3704	323.4675	917.2295
6	356.0276	-41.91201	-32.11602	430.4720	323.7759	900.0293
7	355.9964	-41.84261	-32.02402	430.0556	323.9115	892.2330
8	355.9817	-41.81050	-31.98178	429.8630	323.9724	888.6705
9	355.9749	-41.79567	-31.96235	429.7740	323.9999	887.0355
10	355.9717	-41.78882	-31.95339	429.7329	324.0125	886.2831
11	355.9703	-41.78566	-31.94928	429.7139	324.0183	885.9361
12	355.9696	-41.78422	-31.94739	429.7052	324.0210	885.7762
13	355.9693	-41.78355	-31.94651	429.7012	324.0222	885.7018
14	355.9691	-41.78322	-31.94608	429.6993	324.0228	885.6691
15	355.9691	-41.78308	-31.94590	429.6984	324.0230	885.6527
16	355.9690	-41.78302	-31.94582	429.6980	324.0232	885.6473
17	355.9690	-41.78298	-31.94578	429.6979	324.0232	885.6418
18	355.9690	-41.78298	-31.94577	429.6978	324.0233	885.6418
19	355.9690	-41.78298	-31.94577	429.6978	324.0233	885.6418

The final estimates, of course, coincide with the estimates found by direct least squares, which are obtained upon solving the normal equations

$$\bar{y} = \beta_0 + \frac{\beta_1}{7}; \quad y_{+(1)} - y_{-(1)} = \beta_0 + 5\beta_1 - \beta_2; \quad y_{+(2)} - y_{-(2)} = \beta_1 + 4\beta_2.$$

2.3.3 Example 2.3: Missing Values in a Latin Square Design

Consider the data in Table 2.4 (adapted from Cochran and Cox, 1957) on errors in shoot heights (of wheat in centimeters) of six samplers (treatments) arranged in a Latin square, the columns being six areas and the rows being the order in which the areas were sampled. Two of the observations (indicated by ??) are missing.

Table 2.4 Sampler's Errors in Shoot Heights (cm) (6×6 Latin Square).

Order	Areas					
	1	2	3	4	5	6
I	F 3.5	B 4.2	A 6.7	D 6.6	C 4.1	E 3.8
II	B 8.9	F 1.9	D ??	A 4.5	E 2.4	C 5.8
III	C 9.6	E 3.7	F-2.7	B 3.7	D 6.0	A 7.0
IV	D 10.5	C 10.2	B 4.6	E 3.7	A 5.1	F 3.8
V	E ??	A 7.2	C 4.0	F-3.3	B 3.5	D 5.0
VI	A 5.9	D 7.6	E-0.7	C 3.0	F 4.0	B 8.6

Source: Adapted from Cochran and Cox (1957)

Assuming the standard additive model of row, column, and treatment effects, in a complete Latin square, the estimates of the row, column, and treatment differences are simply those obtained from the corresponding means. Thus the expected value of the observation in row i and column j with treatment k is $(3R_i + 3C_j + 3T_k - G)/18$, where R_i , C_j , and T_k are the totals of the i th row, j th column, and the k th treatment, respectively, and G is the total of all the observations. Thus, this formula will be used for predicting a missing value in the Healy-Westmacott procedure. The residual sum of squares is computed by the standard analysis of variance resolving the total sum of squares into its components of row, column, treatment, and residual sums of squares. In Table 2.5, we present the results obtained by the Healy-Westmacott procedure started from initial estimates based on the effect means calculated from the available data.

Table 2.5 Results of the Healy-Westmacott Procedure on the Data of Example 2.3.

Iteration	\hat{y}_{23}	\hat{y}_{51}	Residual SS
1	4.73	3.598	65.847
2	4.73	3.598	65.847

The final estimates of row, column, and treatment totals are:

Row:	28.9,	28.23,	27.3,	37.9,	20.91,	28.4
Column:	42.0,	34.8,	16.63,	18.2,	25.1,	34.0
Treatment:	36.4,	33.5,	36.7,	40.43,	16.5,	7.2

2.3.4 Healy-Westmacott Procedure as an EM Algorithm

The Healy-Westmacott procedure can be looked upon as an application of the EM algorithm. Suppose we assume that the conditional distribution of the j th response variable y_j , given the vector x_j , is normal with mean $\beta^T x_j$, where x_j is a row of the design matrix X , and

variance σ^2 (not depending upon \mathbf{X}). Then the least-squares estimate of β and the estimate obtained in the least-squares analysis for σ^2 are also the MLE's. Hence the problem of estimation of the parameter vector,

$$\Psi = (\beta, \sigma^2)^T,$$

can be looked upon as a ML estimation problem, and thus in the case of missing values, the EM algorithm can be profitably applied, if the complete-data problem is simpler. Let us consider the EM algorithm in this role. Suppose that the data are labeled so that the first m observations are complete with the last $(n - m)$ cases having the response y_j missing. Note that we are using here the universal notation \mathbf{X} to denote the design matrix. Elsewhere in this book, \mathbf{X} denotes the random vector corresponding to the complete-data vector \mathbf{x} . Also, the use in this section of \mathbf{x}_j as the j th row of \mathbf{X} is not to be confused as being a replication of \mathbf{x} , which is used elsewhere in the book to denote the complete-data vector.

The incomplete-data log likelihood function is given by

$$\begin{aligned} \log L(\Psi) &= -\frac{1}{2}m \log(2\pi) - \frac{1}{2}m \log \sigma^2 - \frac{1}{2} \sum_{j=1}^m y_j^2 / \sigma^2 \\ &\quad - \frac{1}{2} \sum_{j=1}^m (\beta^T \mathbf{x}_j)^2 / \sigma^2 + \sum_{j=1}^m y_j (\beta^T \mathbf{x}_j) / \sigma^2. \end{aligned} \quad (2.16)$$

The complete-data log likelihood is the same except that the summation in (2.16) ranges up to n . Concerning the E-step on the $(k + 1)$ th iteration, it follows from (2.16) that in order to compute the current conditional expectation of the complete-data log likelihood function, we require the conditional expectations of the first two moments of the missing responses Y_j given the vector

$$\mathbf{y} = (y_1, \dots, y_m)^T$$

of the observed responses and the design matrix \mathbf{X} . They are given by

$$E_{\Psi^{(k)}}(Y_j | \mathbf{y}, \mathbf{X}) = \mathbf{x}_j^T \beta^{(k)}$$

and

$$E_{\Psi^{(k)}}(Y_j^2 | \mathbf{y}, \mathbf{X}) = (\mathbf{x}_j^T \beta^{(k)})^2 + \sigma^{(k)2} \quad (2.17)$$

for $j = m + 1, \dots, n$.

Implementation of the M-step at the $(k + 1)$ th iteration is straightforward, since we use the least-squares computation of the complete-data design to find $\beta^{(k+1)}$, and $\sigma^{(k+1)2}$ is given by the current residual mean square. We thus obtain

$$\beta^{(k+1)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}^{(k)}$$

and

$$\sigma^{(k+1)2} = \frac{1}{n} \left\{ \sum_{j=1}^m (y_j - \beta^{(k)T} \mathbf{x}_j)^2 + (n - m) \sigma^{(k)2} \right\},$$

where

$$\mathbf{Y}^{(k)} = (y_1, \dots, y_m, y_{m+1}^{(k)}, \dots, y_n^{(k)})^T,$$

and

$$y_j^{(k)} = \mathbf{x}_j^T \beta^{(k)}.$$

The first equation does not involve $\sigma^{(k)^2}$, and at convergence

$$\sigma^{(k+1)^2} = \sigma^{(k)^2} = \hat{\sigma}^2.$$

Thus

$$\hat{\sigma}^2 = \frac{1}{n} \left\{ \sum_{j=1}^m (y_j - \hat{\beta}^T \mathbf{x}_j)^2 + (n-m)\hat{\sigma}^2 \right\},$$

or

$$\hat{\sigma}^2 = \frac{1}{m} \sum_{j=1}^m (y_j - \hat{\beta}^T \mathbf{x}_j)^2.$$

Hence the EM algorithm can omit the M-step for $\sigma^{(k)^2}$ and the calculation of (2.17) on the E-step, and compute $\hat{\sigma}^2$ after the convergence of the sequence $\{\beta^{(k)}\}$. It can be easily seen that this is exactly the Healy–Westmacott procedure described above, confirming that the procedure is an example of the EM algorithm; see Little and Rubin (2002, pages 164–172) for further discussion.

2.4 EXAMPLE 2.4: MULTINOMIAL WITH COMPLEX CELL STRUCTURE

In Section 1.4.2, we gave an introductory example on the EM algorithm applied to a multinomial distribution with cell probabilities depending on a single unknown parameter. The EM algorithm can handle more complicated problems of this type that occur in genetics. To illustrate this, we consider a multinomial problem with cell probabilities depending on two unknown parameters.

Suppose we have $n = 435$ observations on a multinomial with four cells with cell probability structure given in Table 2.6. Also given in the table are observed frequencies in these cells.

Table 2.6 Observed Multinomial Data for Example 2.4.

Category (Cell)	Cell Probability	Observed Frequency
O	r^2	$n_O = 176$
A	$p^2 + 2pr$	$n_A = 182$
B	$q^2 + 2qr$	$n_B = 60$
AB	$2pq$	$n_{AB} = 17$

The observed data are therefore given by the vector of cell frequencies

$$\mathbf{y} = (n_O, n_A, n_B, n_{AB})^T.$$

The vector of unknown parameters is

$$\Psi = (p, q)^T,$$

since $r = 1 - p - q$. The object is to find the MLE of Ψ on the basis of \mathbf{y} . This is a well-known problem of gene frequency estimation in genetics, and is discussed, for example, in Kempthorne (1957), Elandt-Johnson (1971), and Rao (1973).

The log likelihood function for Ψ , apart from an additive constant, is

$$\log L(\Psi) = 2n_0 \log r + n_A \log(p^2 + 2pr) + n_B \log(q^2 + 2qr) + n_{AB} \log(2pq),$$

which does not admit a closed-form solution for $\hat{\Psi}$, the MLE of Ψ .

Let us denote the cell frequencies by π_j ($j = 1, 2, 3, 4$). Then their first and second derivatives with respect to Ψ are as follows:

$$\begin{aligned}\frac{\partial \pi_1(\Psi)}{\partial \Psi} &= \begin{pmatrix} 2r \\ 2r \end{pmatrix}; & \frac{\partial \pi_2(\Psi)}{\partial \Psi} &= \begin{pmatrix} 2r \\ -2p \end{pmatrix} \\ \frac{\partial \pi_3(\Psi)}{\partial \Psi} &= \begin{pmatrix} -2q \\ 2r \end{pmatrix}; & \frac{\partial \pi_4(\Psi)}{\partial \Psi} &= \begin{pmatrix} 2q \\ 2p \end{pmatrix}\end{aligned}\quad (2.18)$$

$$\begin{aligned}\frac{\partial^2 \pi_1(\Psi)}{\partial \Psi \partial \Psi^T} &= \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}; & \frac{\partial^2 \pi_2(\Psi)}{\partial \Psi \partial \Psi^T} &= \begin{pmatrix} -2 & -2 \\ -2 & 0 \end{pmatrix} \\ \frac{\partial^2 \pi_3(\Psi)}{\partial \Psi \partial \Psi^T} &= \begin{pmatrix} 0 & -2 \\ -2 & -2 \end{pmatrix}; & \frac{\partial^2 \pi_4(\Psi)}{\partial \Psi \partial \Psi^T} &= \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}.\end{aligned}\quad (2.19)$$

This leads to the likelihood equation as

$$\partial \log L(\Psi) / \partial \Psi = \sum_{j=1}^4 \left(\frac{n_j}{\pi_j} \right) \frac{\partial \pi_j(\Psi)}{\partial \Psi} = \mathbf{0}, \quad (2.20)$$

and the Hessian of the log likelihood as

$$\partial^2 \log L(\Psi) / \partial \Psi \partial \Psi^T = \sum_{j=1}^4 n_j \left\{ \left(\frac{1}{\pi_j} \right) \frac{\partial^2 \pi_j(\Psi)}{\partial \Psi \Psi^T} - \left(\frac{1}{\pi_j^2} \right) \frac{\partial \pi_j(\Psi)}{\partial \Psi} \left(\frac{\partial \pi_j(\Psi)}{\partial \Psi^T} \right) \right\}. \quad (2.21)$$

The Fisher (expected) information matrix is given by

$$\begin{aligned}\mathcal{I}(\Psi) &= E\{-\partial^2 \log L(\Psi) / \partial \Psi \partial \Psi^T\} \\ &= n \left\{ \sum_{j=1}^4 \left(\frac{1}{\pi_j} \right) \frac{\partial \pi_j(\Psi)}{\partial \Psi} \left(\frac{\partial \pi_j(\Psi)}{\partial \Psi^T} \right) \right\},\end{aligned}\quad (2.22)$$

yielding at $\Psi = \hat{\Psi}$ the following covariance matrix of the estimates

$$\begin{pmatrix} 0.000011008 & -0.000103688 \\ -0.000103688 & 0.000040212 \end{pmatrix}; \quad (2.23)$$

see Monahan (2001) for an interesting discussion of Newton's method, Scoring method, and the EM algorithm for this example.

Let us now discuss the application of the EM algorithm to this problem. In applying the EM algorithm to this problem, a natural choice for the complete-data vector is

$$\mathbf{x} = (n_O, \mathbf{z}^T)^T, \quad (2.24)$$

where

$$\mathbf{z} = (n_{AA}, n_{AO}, n_{BB}, n_{BO})^T$$

Table 2.7 Complete-Data Structure for Example 2.4.

Category (Cell)	Cell Probability	Notation for Frequency
O	r^2	n_O
AA	p^2	n_{AA}
AO	$2pr$	n_{AO}
BB	q^2	n_{BB}
BO	$2qr$	n_{BO}
AB	$2pq$	n_{AB}

represents the unobservable or ‘missing’ data, taken to be the frequencies n_{AA} , n_{AO} , n_{BB} , and n_{BO} , corresponding to the middle cells as given in Table 2.7. Notice that since the total frequency n is fixed, the five cell frequencies in (2.24) are sufficient to represent the complete data. If we take the distribution of \boldsymbol{x} to be multinomial with n draws with respect to the six cells with probabilities as specified in Table 2.7, then it is obvious that the vector \boldsymbol{y} of observed frequencies has the required multinomial distribution, as specified in Table 2.6.

The complete-data log likelihood function for Ψ can be written in the form (apart from an additive constant) as

$$\log L_c(\Psi) = 2n_A^+ \log p + 2n_B^+ \log q + 2n_O^+ \log r, \quad (2.25)$$

where

$$n_A^+ = n_{AA} + \frac{1}{2}n_{AO} + \frac{1}{2}n_{AB},$$

$$n_B^+ = n_{BB} + \frac{1}{2}n_{BO} + \frac{1}{2}n_{AB},$$

and

$$n_O^+ = n_O + \frac{1}{2}n_{AO} + \frac{1}{2}n_{BO}.$$

It can be seen that (2.25) has the form of a multinomial log likelihood for the frequencies $2n_A^+$, $2n_B^+$, and $2n_O^+$ with respect to three cells with probabilities p , q , and r , respectively. Thus the complete-data MLE’s of these probabilities obtained by maximization of (2.25) are given by

$$\hat{p} = \frac{n_A^+}{n}; \quad \hat{q} = \frac{n_B^+}{n}. \quad (2.26)$$

As seen in Section 1.5.3, the E- and M-steps simplify when the complete-data likelihood function belongs to the regular exponential family, as in this example. The E-step requires just the calculation of the current conditional expectation of the sufficient statistic for Ψ , which here is $(n_A^+, n_B^+)^T$. The M-step then obtains $\Psi^{(k+1)}$ as a solution of the equation obtained by equating this expectation to Ψ . In effect for this problem, $\Psi^{(k+1)}$ is given by replacing n_A^+ and n_B^+ in the right-hand side of (2.26) by their current conditional expectations given the observed data.

To compute these conditional expectations of n_A^+ and n_B^+ (the E-step), we require the conditional expectation of the unobservable data \boldsymbol{z} for this problem. Consider the first element of \boldsymbol{z} , which from (2.24) is n_{AA} . Then it is easy to verify that conditional on

\mathbf{y} , effectively n_A , n_{AA} has a binomial distribution with sample size n_A and probability parameter

$$p^{(k)^2} / (p^{(k)^2} + 2p^{(k)}r^{(k)}),$$

with $\Psi^{(k)}$ used in place of the unknown parameter vector Ψ on the $(k+1)$ th iteration. Thus the current conditional expectation of n_{AA} given \mathbf{y} is obtained by

$$E_{\Psi^{(k)}}(n_{AA}) = n_{AA}^{(k)},$$

where

$$n_{AA}^{(k)} = n_A p^{(k)^2} / (p^{(k)^2} + 2p^{(k)}r^{(k)}). \quad (2.27)$$

Similarly, the current conditional expectations of n_{AO} , n_{BB} , and n_{BO} , given \mathbf{y} , can be calculated.

Execution of the M-step gives

$$p^{(k+1)} = (n_{AA}^{(k)} + \frac{1}{2}n_{AO}^{(k)} + \frac{1}{2}n_{AB}^{(k)})/n \quad (2.28)$$

and

$$q^{(k+1)} = (n_{BB}^{(k)} + \frac{1}{2}n_{BO}^{(k)} + \frac{1}{2}n_{AB}^{(k)})/n. \quad (2.29)$$

Results of the EM algorithm for this problem are given in Table 2.8. The MLE of Ψ can be taken to be the value of $\Psi^{(k)}$ on iteration $k = 4$. Although not formulated in the language of the EM algorithm, this idea was current in the genetics literature as the gene-counting method (see Ceppellini et al., 1957).

Table 2.8 Results of the EM Algorithm for Example 2.4.

Iteration	$p^{(k)}$	$q^{(k)}$	$r^{(k)}$	$-\log L(\Psi^{(k)})$
0	0.26399	0.09299	0.64302	2.5619001
1	0.26436	0.09316	0.64248	2.5577875
2	0.26443	0.09317	0.64240	2.5577729
3	0.26444	0.09317	0.64239	2.5577726
4	0.26444	0.09317	0.64239	2.5577726

2.5 EXAMPLE 2.5: ANALYSIS OF PET AND SPECT DATA

The EM algorithm has been employed in ML estimation of parameters in a computerized image reconstruction process such as SPECT (single-photon emission computed tomography) or PET (positron emission tomography). An excellent account of the statistical aspects of emission tomography may be found in the papers by Kay (1994) and McColl, Holmes, and Ford (1994), which are contained in the special issue of the journal *Statistical Methods in Medical Research* on the topic of emission tomography. In both PET and SPECT a radioactive tracer is introduced into the organ under study of the human or animal patient. The radioisotope is incorporated into a molecule that is absorbed into the tissue of the organ and resides there in concentrations that indicate levels of metabolic activity and blood flow. As the isotope decays, it emits either single photons (SPECT) or positrons (PET), which are counted by bands of gamma detectors strategically placed around the patient's body.

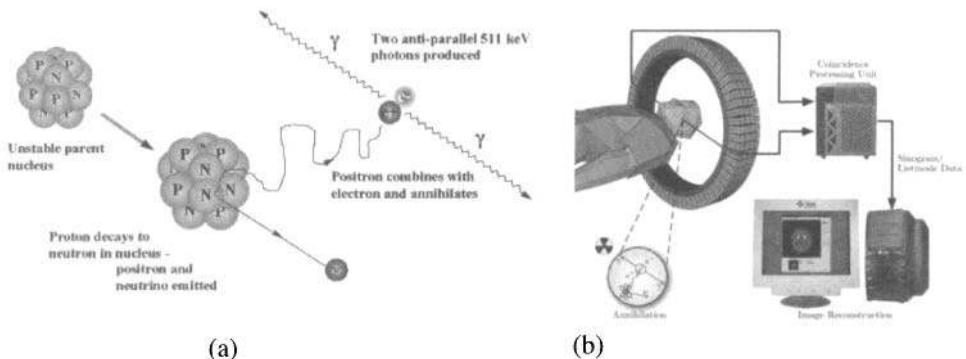


Figure 2.1 (a) Positron annihilation. From Badawi (1999); (b) PET detector. From Wikipedia.

The method of collection of the data counts differs quite considerably for PET and SPECT because of the fundamental differences in their underlying physical processes. SPECT isotopes are gamma emitters that tend to have long half lives. PET isotopes emit positrons which annihilate with nearby electrons in the tissue to generate pairs of photons that fly off on paths almost 180° apart. In both imaging modalities, not all emissions are counted because the photons may travel along directions that do not cross the detectors or because they may be attenuated by the body's tissues.

With PET and SPECT, the aim is to estimate the spatial distribution of the isotope concentration in the organ on the basis of the projected counts recorded at the detectors. In a statistical framework, it is assumed that the emissions occur according to a spatial Poisson point process in the region under study with an unknown intensity function, which is usually referred to as the emission density. In order to estimate the latter, the process is discretized as follows. The space over which the reconstruction is required is finely divided into a number n of rectangular pixels (or voxels in three-dimensions), and it is assumed that the (unknown) emission density is a constant λ_i for the i th pixel ($i = 1, \dots, n$). Let y_j denote the number of counts recorded by the j th detector ($j = 1, \dots, d$), where d denotes the number of detectors. As it is common to work with arrays of 128×128 or 256×256 square pixels, the detectors move within the plane of measurement during a scan in order to record more observations than parameters. They count for only a short time at each position. Figure 2.1 describes positron annihilation and the schema of a PET detector.

After this discretization of the problem, reconstruction aims to infer the vector of emission densities

$$\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)^T$$

from the vector of the observable counts $\mathbf{y} = (y_1, \dots, y_d)^T$. In both PET and SPECT, it leads to a Poisson regression model for the counts. Given the vector $\boldsymbol{\lambda}$ of isotope densities, the counts y_1, \dots, y_d , are conditionally independent according to a Poisson distribution, namely

$$Y_j \sim P(\mu_j), \quad (2.30)$$

where the mean μ_j of Y_j is modeled as

$$\mu_j = \sum_{i=1}^n \lambda_i p_{ij} \quad (j = 1, \dots, d), \quad (2.31)$$

and p_{ij} is the conditional probability that a photon/positron is counted by the j th detector given that it was emitted from within the i th pixel.

The advantage of this statistical approach is that it takes into account the Poisson variation underlying the data, unlike the class of deterministic methods based on a filtered back-projection approach. The conditional probabilities p_{ij} , which appear as the coefficients in the additive Poisson regression model (2.31), depend on the geometry of the detection system, the activity of the isotope and exposure time, and the extent of attenuation and scattering between source and detector. The specification of the p_{ij} is much more complicated with SPECT than PET since attenuation can be neglected for the latter; see, for example, Green (1990b), McColl et al. (1994), and Kay (1994).

We now outline the application of the EM algorithm to this problem of tomography reconstruction. This approach was pioneered independently by Shepp and Vardi (1982) and Lange and Carson (1984), and developed further by Vardi, Shepp, and Kaufman (1985).

An obvious choice of the complete-data vector in this example is $\mathbf{x} = (\mathbf{y}^T, \mathbf{z}^T)^T$, where the vector \mathbf{z} consists of the unobservable data counts z_{ij} with z_{ij} defined to be the number of photons/positrons emitted within pixel i and recorded at the j th detector ($i = 1, \dots, n$; $j = 1, \dots, d$). It is assumed that given $\boldsymbol{\lambda}$, the $\{Z_{ij}\}$ are conditionally independent, with each Z_{ij} having a Poisson distribution specified as

$$Z_{ij} \sim P(\lambda_i p_{ij}) \quad (i = 1, \dots, d; j = 1, \dots, n).$$

Since

$$y_j = \sum_{i=1}^n z_{ij}, \quad (j = 1, \dots, d),$$

it is obvious that these assumptions for the unobservable data $\{z_{ij}\}$ imply the model (2.30) for the incomplete data $\{y_j\}$ in \mathbf{y} .

The complete-data log likelihood is given by

$$\log L_c(\boldsymbol{\lambda}) = \sum_{i=1}^n \sum_{j=1}^d \{-\lambda_i p_{ij} + z_{ij} \log(\lambda_i p_{ij}) - \log z_{ij}!\}. \quad (2.32)$$

As the Poisson distribution belongs to the linear exponential family, (2.32) is linear in the unobservable data z_{ij} . Hence the E-step (on the $(k+1)$ th iteration) simply requires the calculation of the conditional expectation of Z_{ij} given the observed data \mathbf{y} , using the current fit $\boldsymbol{\lambda}^{(k)}$ for $\boldsymbol{\lambda}$. From a standard probability result, the conditional distribution of Z_{ij} given \mathbf{y} and $\boldsymbol{\lambda}^{(k)}$ is binomial with sample size parameter y_j and probability parameter

$$\lambda_i p_{ij} / \sum_{h=1}^n \lambda_h p_{hj} \quad (i = 1, \dots, n; j = 1, \dots, d). \quad (2.33)$$

Using (2.33), it follows that

$$E_{\boldsymbol{\lambda}^{(k)}}(Z_{ij} | \mathbf{y}) = z_{ij}^{(k)},$$

where

$$z_{ij}^{(k)} = y_j \lambda_i^{(k)} p_{ij} / \sum_{h=1}^n \lambda_h^{(k)} p_{hj}. \quad (2.34)$$

With z_{ij} replaced by $z_{ij}^{(k)}$ in (2.32), application of the M-step on the $(k + 1)$ th iteration gives

$$\begin{aligned} \lambda_i^{(k+1)} &= q_i^{-1} \sum_{j=1}^d p_{ij} E_{\lambda_i^{(k)}}(Z_{ij} | \mathbf{y}) \\ &= \lambda_i^{(k)} q_i^{-1} \sum_{j=1}^d \left\{ y_j p_{ij} / \sum_{h=1}^n \lambda_h^{(k)} p_{hj} \right\} \quad (i = 1, \dots, n), \end{aligned} \quad (2.35)$$

where

$$q_i = \sum_{j=1}^d p_{ij}$$

is the probability that an emission from the i th pixel is recorded by one of the d detectors ($i = 1, \dots, n$). As noted by Vardi et al. (1985), it is much simpler and involves no loss of generality to take $q_i = 1$. This is because we can work with the density of emission counts that are actually detected; that is, we let λ_i represent the mean number of emissions from the i th pixel that are detected.

It has been shown (Vardi et al., 1985) that $\lambda^{(k)}$ converges to a global maximizer of the incomplete-data likelihood function $L(\lambda)$, but it will not be unique if the number of pixels n is greater than the number of detectors d ; see Byrne (1993). An attractive feature of (2.35) is that the positivity constraints on the estimates of the λ_i are satisfied, providing the initial estimates $\lambda_i^{(0)}$ are all positive.

After a certain number of iterations, the EM algorithm produces images that begin to deteriorate. The point at which deterioration begins depends on the number of the projected counts y_j ($j = 1, \dots, d$). The more the counts the later it begins. The reason for this deterioration in the reconstructed image is that for a low number of counts increasing the value of the likelihood function increases the distortion due to noise. As a consequence, various modifications to the EM algorithm have been proposed to produce more useful final images. Although these modified versions of the EM algorithm are algorithms that can be applied to a variety of problems, the motivation for their development in the first instance came in the context of improving the quality of images reconstructed from PET and SPECT data. Hence it is in this context that we shall outline in Chapter 5, some of these modifications of the EM algorithm; for a review and a bibliography of the EM algorithm in tomography, see Krishnan (1995). Further modifications of the EM algorithm in its application to this problem have been given by Fessler and Hero (1994) and Meng and van Dyk (1997).

Richardson (1972) and Lucy (1974) independently obtained the same algorithm in the context of restoration of astronomical images and to astronomers and astrophysicists it is known as the Richardson-Lucy Algorithm or Richardson-Lucy Deconvolution. There are many instances of its application in astronomy, one of which is by Molina, Núñez, Cortijo, and Mateos (2001) for restoration of the image of Saturn as in Figure 2.2.

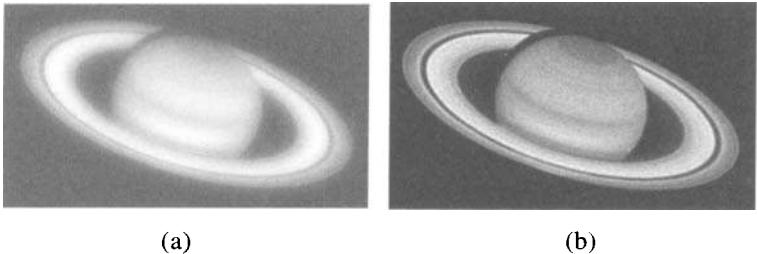


Figure 2.2 (a) Blurred Image of Saturn obtained with WF/PC camera of Hubble Space Telescope reproduced with permission from Space Telescope Science Institute (STSCI); (b) its Reconstructed Image using the Richardson-Lucy Algorithm from Molina et al. (2001) reproduced with permission from IEEE ©IEEE.

2.6 EXAMPLE 2.6: MULTIVARIATE t -DISTRIBUTION WITH KNOWN DEGREES OF FREEDOM

2.6.1 ML Estimation of Multivariate t -Distribution

The multivariate t -distribution has many potential applications in applied statistics; see Kotz and Nadarajah (2004). A p -dimensional random variable \mathbf{W} is said to have a multivariate t -distribution $t_p(\boldsymbol{\mu}, \Sigma, \nu)$ with location $\boldsymbol{\mu}$, positive definite inner product matrix Σ , and ν degrees of freedom if given the weight u ,

$$\mathbf{W} | u \sim N(\boldsymbol{\mu}, \Sigma/u), \quad (2.36)$$

where the random variable U corresponding to the weight u is distributed as

$$U \sim \text{gamma}(\frac{1}{2}\nu, \frac{1}{2}\nu). \quad (2.37)$$

The gamma (α, β) density function $f(u; \alpha, \beta)$ is given by

$$f(u; \alpha, \beta) = \{\beta^\alpha u^{\alpha-1}/\Gamma(\alpha)\} \exp(-\beta u) I_{[0, \infty)}(u); \quad (\alpha, \beta > 0).$$

On integrating out u from the joint density function of \mathbf{W} and U that can be formed from (2.36) and (2.37), the density function of \mathbf{W} is given by

$$f_p(\mathbf{w}; \boldsymbol{\mu}, \Sigma, \nu) = \frac{\Gamma(\frac{\nu+p}{2}) |\Sigma|^{-1/2}}{(\pi\nu)^{\frac{1}{2}p} \Gamma(\frac{\nu}{2}) \{1 + \delta(\mathbf{w}, \boldsymbol{\mu}; \Sigma)/\nu\}^{\frac{1}{2}(\nu+p)}}, \quad (2.38)$$

where

$$\delta(\mathbf{w}, \boldsymbol{\mu}; \Sigma) = (\mathbf{w} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{w} - \boldsymbol{\mu})$$

denotes the Mahalanobis squared distance between \mathbf{w} and $\boldsymbol{\mu}$ (with Σ as the covariance matrix). As ν tends to infinity, U converges to one with probability one, and so \mathbf{W} becomes marginally multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ .

We consider now the application of the EM algorithm for ML estimation of the parameters $\boldsymbol{\mu}$ and Σ in the t -density (2.38) in the case where the degrees of freedom ν is known. For

instance, ν can be assumed to be known in statistical analyses where different specified degrees of freedom ν are used for judging the robustness of the analyses; see Lange, Little, and Taylor (1989) and Lange and Sinsheimer (1993). For $\nu < \infty$, ML estimation of μ is robust in the sense that observations with large Mahalanobis distances are downweighted. This can be clearly seen from the form of the equation (2.48) to be derived for the MLE of μ . As ν decreases, the degree of downweighting of outliers increases.

The EM algorithm for known ν is given in Rubin (1983), and is extended to the case with missing data in \mathbf{W} in Little (1988) and in Little and Rubin (1987, 2002, pages 257–264). Liu and Rubin (1994, 1995) and Little and Rubin (2002, pages 183–184) have shown how the MLE can be found much more efficiently by using the ECME algorithm. The use of the latter algorithm for this problem is to be described in Chapter 5, where the general case of unknown ν is to be considered. As cautioned by Liu and Rubin (1995), care must be taken especially with small or unknown ν , because the likelihood function can have many spikes with very high likelihood values but little associated posterior mass under any reasonable prior. In that case, the associated parameter estimates may be of limited practical interest by themselves, even though formally they are local or even global maxima of the likelihood function. It is, nevertheless, important to locate such maxima because they can critically influence the behavior of iterative simulation algorithms designed to summarize the entire posterior distribution; see Gelman and Rubin (1992).

Suppose that $\mathbf{w}_1, \dots, \mathbf{w}_n$ denote an observed random sample from the $t_p(\mu, \Sigma, \nu)$ distribution. That is,

$$\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T$$

denotes the observed data vector. The problem is to find the MLE of Ψ on the basis of \mathbf{y} , where Ψ contains the elements of μ and the distinct elements of Σ . From (2.38), the log likelihood function for Ψ that can be formed from \mathbf{y} is

$$\begin{aligned}\log L(\Psi) &= \sum_{j=1}^n \log f_p(\mathbf{w}_j; \mu, \Sigma, \nu), \\ &= -\frac{1}{2}np \log(\pi\nu) + n\{\log \Gamma(\frac{\nu+p}{2}) - \log \Gamma(\frac{1}{2}\nu)\} \\ &\quad -\frac{1}{2}n \log |\Sigma| + \frac{1}{2}n(\nu+p) \log \nu \\ &\quad -\frac{1}{2}(\nu+p) \sum_{j=1}^n \log\{\nu + \delta(\mathbf{w}_j; \mu, \Sigma)\},\end{aligned}\tag{2.39}$$

which does not admit a closed form solution for the MLE of Ψ .

In the light of the definition (2.36) of this t -distribution, it is convenient to view the observed data \mathbf{y} as incomplete. The complete-data vector \mathbf{x} is taken to be

$$\mathbf{x} = (\mathbf{y}^T, \mathbf{z}^T)^T,$$

where

$$\mathbf{z} = (u_1, \dots, u_n)^T.$$

The missing variables u_1, \dots, u_n are defined so that

$$\mathbf{W}_j \mid u_j \sim N(\mu, \Sigma/u_j),\tag{2.40}$$

independently for $j = 1, \dots, n$, and

$$U_1, \dots, U_n \stackrel{\text{iid}}{\sim} \text{gamma}(\frac{1}{2}\nu, \frac{1}{2}\nu).\tag{2.41}$$

Here in this example, the missing data vector \mathbf{z} consists of variables that would never be observable as data in the usual sense.

Because of the conditional structure of the complete-data model specified by (2.36) and (2.37), the complete-data likelihood function can be factored into the product of the conditional density of \mathbf{W} given \mathbf{z} and the marginal density of \mathbf{Z} . Accordingly, the complete-data log likelihood can be written as

$$\log L_c(\boldsymbol{\Psi}) = \log L_{1c}(\boldsymbol{\Psi}) + a(\mathbf{z}),$$

where

$$\begin{aligned}\log L_{1c}(\boldsymbol{\Psi}) &= -\frac{1}{2}np\log(2\pi) - \frac{1}{2}n\log|\boldsymbol{\Sigma}| \\ &\quad -\frac{1}{2}\sum_{j=1}^n u_j(\mathbf{w}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{w}_j - \boldsymbol{\mu})\end{aligned}\quad (2.42)$$

and

$$\begin{aligned}a(\mathbf{z}) &= -n\log\Gamma(\frac{1}{2}\nu) + \frac{1}{2}n\nu\log(\frac{1}{2}\nu) \\ &\quad + \frac{1}{2}\nu\sum_{j=1}^n (\log u_j - u_j) - \sum_{j=1}^n \log u_j.\end{aligned}\quad (2.43)$$

Now the E-step on the $(k+1)$ th iteration of the EM algorithm requires the calculation of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k+1)})$, the current conditional expectation of the complete-data log likelihood function $\log L_c(\boldsymbol{\Psi})$. In the present case of known ν , we need focus only on the first term $\log L_{1c}(\boldsymbol{\Psi})$ in the expression (2.42) for $\log L_c(\boldsymbol{\Psi})$ (since the other term does not involve unknown parameters). As this term is linear in the unobservable data u_j , the E-step is effected simply by replacing u_j with its current conditional expectation given w_j .

Since the gamma distribution is the conjugate prior distribution for U , it is not difficult to show that the conditional distribution of U given $\mathbf{W} = \mathbf{w}$ is

$$U \mid \mathbf{w} \sim \text{gamma}(m_1, m_2), \quad (2.44)$$

where

$$m_1 = \frac{1}{2}(\nu + p)$$

and

$$m_2 = \frac{1}{2}\{\nu + \delta(\mathbf{w}, \boldsymbol{\mu}; \boldsymbol{\Sigma})\}. \quad (2.45)$$

From (2.44), we have that

$$E(U \mid \mathbf{w}) = \frac{\nu + p}{\nu + \delta(\mathbf{w}, \boldsymbol{\mu}; \boldsymbol{\Sigma})}. \quad (2.46)$$

Thus from (2.46),

$$E_{\boldsymbol{\Psi}^{(k)}}(U_j \mid \mathbf{w}_j) = u_j^{(k)},$$

where

$$u_j^{(k)} = \frac{\nu + p}{\nu + \delta(\mathbf{w}_j, \boldsymbol{\mu}^{(k)}; \boldsymbol{\Sigma}^{(k)})}. \quad (2.47)$$

The M-step is easily implemented on noting that $L_{1c}(\boldsymbol{\Psi})$ corresponds to the likelihood function formed from n independent observations $\mathbf{w}_1, \dots, \mathbf{w}_n$ with common mean $\boldsymbol{\mu}$ and

covariance matrices $\Sigma/u_1, \dots, \Sigma/u_n$, respectively. After execution of the E-step, each u_j is replaced by $u_j^{(k)}$, and so the M-step is equivalent to computing the weighted sample mean and sample covariance matrix of $\mathbf{w}_1, \dots, \mathbf{w}_n$ with weights $u_1^{(k)}, \dots, u_n^{(k)}$. Hence

$$\boldsymbol{\mu}^{(k+1)} = \sum_{j=1}^n u_j^{(k)} \mathbf{w}_j / \sum_{j=1}^n u_j^{(k)} \quad (2.48)$$

and

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n} \sum_{j=1}^n u_j^{(k)} (\mathbf{w}_j - \boldsymbol{\mu}^{(k+1)}) (\mathbf{w}_j - \boldsymbol{\mu}^{(k+1)})^T. \quad (2.49)$$

It can be seen in this case of known ν that the EM algorithm is equivalent to iteratively reweighted least squares. The E-step updates the weights $u_j^{(k)}$, while the M-step effectively chooses $\boldsymbol{\mu}^{(k+1)}$ and $\boldsymbol{\Sigma}^{(k+1)}$ by weighted least-squares estimation.

Liu and Rubin (1995) note that the above results can be easily extended to linear models where the mean $\boldsymbol{\mu}$ of \mathbf{W}_j is replaced by $\mathbf{X}_j\beta$, where \mathbf{X}_j is a matrix containing the observed values of some covariates associated with \mathbf{W}_j .

In Sections 5.12.2 and 5.15, we discuss the replacement of the divisor n in (2.49) by $\sum_{j=1}^k u_j^{(k)}$ to improve the speed of convergence.

2.6.2 Numerical Example: Stack Loss Data

As a numerical example of the ML fitting of the t -distribution via the EM algorithm, we present an example from Lange et al. (1989), who analyzed the stack-loss data set of Brownlee (1965), which has been subjected to many robust analyses by various authors. Table 2.9 contains the data and Table 2.10 shows the slope of the regression of stack loss (y_j) on air flow (x_{1j}), temperature (x_{2j}), and acid (x_{3j}) for the linear regression model

$$y_j = \beta_0 + \sum_{i=1}^3 \beta_i x_{ij} + e_j,$$

with t -distributed errors e_j for values of the degrees of freedom ν ranging from $\nu = 0.5$ to $\nu = \infty$ (normal). Also included in Table 2.10 are the values of four other estimators from Ruppert and Carroll (1980) of which two are trimmed least-squares ($\hat{\Psi}_{KB}$, $\hat{\Psi}_{PE}$), and two are the M -estimates of Huber (1964) and Andrews (1974). Also, the results are given for the ML estimate of ν , which was found to be $\hat{\nu} = 1.1$. The values of the log likelihood, $\log L(\hat{\Psi})$, at the solutions are given in the second column for various values of ν . Twice the difference between the best fitting t and normal model log likelihoods is 5.44 which, on reference to the chi-squared distribution with one degree of freedom, suggests asymptotically a significant improvement in fit. As noted by Lange et al. (1989), the results of the fit for the ML case of $\hat{\nu} = 1.1$ are similar to those of Andrews (1974), which Ruppert and Carroll (1980) favored based on the closeness of the fit to the bulk of the data.

2.7 FINITE NORMAL MIXTURES

2.7.1 Example 2.7: Univariate Component Densities

We now extend the problem in Example 1.2 of Section 1.4.3 to the situation where the component densities in the mixture model (1.27) are not completely specified.

Table 2.9 Stack Loss Data.

Air flow	Temperature	Acid	Stack loss	Air flow	Temperature	Acid	Stack loss
x_1	x_2	x_3	y	x_1	x_2	x_3	y
80	27	89	42	58	17	89	14
80	27	88	37	58	17	88	13
75	25	90	37	58	18	82	11
62	24	87	28	58	19	93	12
62	22	87	18	50	18	89	8
62	23	87	18	50	18	86	7
62	24	93	19	50	19	72	8
62	24	93	20	50	19	79	8
58	23	87	15	50	20	80	9
58	18	80	14	56	20	82	15
				70	20	91	15

Source: Adapted from Brownlee (1965).

Table 2.10 Estimates of Regression Coefficients with t -distributed Errors and by Other Methods.

Method	log likelihood	Intercept (β_0)	Air Flow (β_1)	Temperature (β_2)	Acid (β_3)
Normal ($t, \nu = \infty$)	-33.0	-39.92	0.72	1.30	-0.15
$t, \nu = 8$	-32.7	-40.71	0.81	0.97	-0.13
$t, \nu = 4$	-32.1	-40.07	0.86	0.75	-0.12
$t, \nu = 3$	-31.8	-39.13	0.85	0.66	-0.10
$t, \nu = 2$	-31.0	-38.12	0.85	0.56	-0.09
$t, \hat{\nu} = 1.1$	-30.3	-38.50	0.85	0.49	-0.07
$t, \nu = 1$	-30.3	-38.62	0.85	0.49	-0.04
$t, \nu = 0.5$	-31.2	-40.82	0.84	0.54	-0.04
Normal minus four outliers		-37.65	0.80	0.58	-0.07
$\hat{\Psi}_{KB}$		-42.83	0.93	0.63	-0.10
$\hat{\Psi}_{PE}$		-40.37	0.72	0.96	-0.07
Huber		-41.00	0.83	0.91	-0.13
Andrews		-37.20	0.82	0.52	-0.07

Source: Adapted from Lange et al. (1989), with permission of the Journal of the American Statistical Association

In this example, the g component densities are taken to be univariate normal with unknown means μ_1, \dots, μ_g and common unknown variance σ^2 . We henceforth write $f_i(w)$ as $f_i(w; \theta_i)$, where $\theta_i = (\mu_i, \sigma^2)^T$ and

$$\boldsymbol{\theta} = (\mu_1, \dots, \mu_g, \sigma^2)^T$$

contains the distinct unknown parameters in these g normal component densities. The vector $\boldsymbol{\Psi}$ containing all the unknown parameters is now

$$\boldsymbol{\Psi} = (\boldsymbol{\theta}^T, \pi_1, \dots, \pi_{g-1})^T.$$

The normal mixture model to be fitted is thus

$$f(w; \Psi) = \sum_{i=1}^g \pi_i f_i(w; \theta_i), \quad (2.50)$$

where

$$\begin{aligned} f_i(w; \theta_i) &= \phi(w; \mu_i, \sigma^2) \\ &= (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}(w - \mu_i)^2/\sigma^2\right\}. \end{aligned}$$

The estimation of Ψ on the basis of y is only meaningful if Ψ is identifiable; that is, distinct values of Ψ determine distinct members of the family

$$\{f(w; \Psi) : \Psi \in \Omega\},$$

where Ω is the specified parameter space. This is true for normal mixtures in that we can determine Ψ up to a permutation of the component labels. For example, for $g = 2$, we cannot distinguish $(\pi_1, \mu_1, \mu_2, \sigma^2)^T$ from $(\pi_2, \mu_2, \mu_1, \sigma^2)^T$, but this lack of identifiability is of no concern in practice, as it can be easily overcome by the imposition of the constraint $\pi_1 \leq \pi_2$; see McLachlan and Basford (1988, Section 1.5), McLachlan and Peel (2000a, Section 1.14). The reader is further referred to Titterington et al. (1985, Section 3.1) for a lucid account of the concept of identifiability for mixtures.

As in the case of Example 1.2 of Section 1.4.3 where only the mixing proportions were unknown, we take the complete-data vector x to be

$$x = (y^T, z^T)^T,$$

where the unobservable vector z is as defined by (1.30). The complete-data log likelihood function for Ψ is given by (1.33), but where now

$$\begin{aligned} C &= \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log f_i(w_j; \theta_i) \\ &= -\frac{1}{2}n \log(2\pi) \\ &\quad -\frac{1}{2} \sum_{i=1}^g \sum_{j=1}^n z_{ij} \{\log \sigma^2 + (w_j - \mu_i)^2/\sigma^2\}. \end{aligned}$$

The E-step is the same as before, requiring the calculation of (1.35). The M-step now requires the computation of not only (1.37), but also the values $\mu_1^{(k+1)}, \dots, \mu_g^{(k+1)}$ and $\sigma^{(k+1)^2}$ that, along with $\pi_1^{(k)}, \dots, \pi_{g-1}^{(k)}$, maximize $Q(\Psi; \Psi^{(k)})$. Now

$$\sum_{j=1}^n z_{ij} w_j / \sum_{j=1}^n z_{ij} \quad (2.51)$$

and

$$\sum_{i=1}^g \sum_{j=1}^n z_{ij} (w_j - \mu)^2 / n \quad (2.52)$$

are the MLE's of μ_i and σ^2 , respectively, if the z_{ij} were observable. As $\log L_c(\Psi)$ is linear in the z_{ij} , it follows that the z_{ij} in (2.51) and (2.52) are replaced by their current

conditional expectations $z_{ij}^{(k)}$, which here are the current estimates $\tau_i(w_j; \Psi^{(k)})$ of the posterior probabilities of membership of the components of the mixture, given by

$$\tau_i(w_j; \Psi^{(k)}) = \pi_i^{(k)} f_i(w_j; \theta_i^{(k)}) / f(w_j; \Psi^{(k)}) \quad (i = 1, \dots, g).$$

This yields

$$\mu_i^{(k+1)} = \sum_{j=1}^n z_{ij}^{(k)} w_j / \sum_{j=1}^n z_{ij}^{(k)} \quad (i = 1, \dots, g) \quad (2.53)$$

and

$$\sigma^{(k+1)^2} = \sum_{i=1}^g \sum_{j=1}^n z_{ij}^{(k)} (w_j - \mu_i^{(k+1)})^2 / n, \quad (2.54)$$

and $\pi_i^{(k+1)}$ is given by (1.37).

The (vector) likelihood equation for $\hat{\mu}_i$ and $\hat{\sigma}^2$ through direct differentiation of the incomplete-data log likelihood function, $\log L(\Psi)$, is quite easily manipulated to show that it is identifiable with the iterative solutions (2.53) and (2.54) produced by application of the EM algorithm. For example, on differentiating (2.50) with respect to μ_i , we obtain that $\hat{\mu}_i$ satisfies the equation

$$\sum_{j=1}^n \hat{\pi}_i \{ f_i(w_j; \hat{\theta}_i) / f(w_j; \hat{\Psi}) \} (w_j - \hat{\mu}_i) = 0 \quad (i = 1, \dots, g),$$

which can be written as

$$\hat{\mu}_i = \sum_{j=1}^n \tau_i(w_j; \hat{\Psi}) w_j / \sum_{j=1}^n \tau_i(w_j; \hat{\Psi}), \quad (2.55)$$

and which is identifiable with (2.53).

2.7.2 Example 2.8: Multivariate Component Densities

The results above for univariate normal component densities easily generalize to the multivariate case. On the M-step at the $(k + 1)$ th iteration, the updated estimates of the i th component mean μ_i and the common covariance matrix Σ are given by

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^n z_{ij}^{(k)} \mathbf{w}_j / \sum_{j=1}^n z_{ij}^{(k)} \quad (i = 1, \dots, g) \quad (2.56)$$

and

$$\boldsymbol{\Sigma}^{(k+1)} = \sum_{i=1}^g \sum_{j=1}^n z_{ij}^{(k)} (\mathbf{w}_j - \boldsymbol{\mu}_i^{(k+1)}) (\mathbf{w}_j - \boldsymbol{\mu}_i^{(k+1)})^T / n. \quad (2.57)$$

The essence of the EM algorithm for this problem is already there in the work of Day (1969).

In the case of normal components with arbitrary covariance matrices, equation (2.57) is replaced by

$$\boldsymbol{\Sigma}_i^{(k+1)} = \sum_{j=1}^n z_{ij}^{(k)} (\mathbf{w}_j - \boldsymbol{\mu}_i^{(k+1)}) (\mathbf{w}_j - \boldsymbol{\mu}_i^{(k+1)})^T / \sum_{j=1}^n z_{ij}^{(k)} \quad (i = 1, \dots, g). \quad (2.58)$$

The likelihood function $L(\Psi)$ tends to have multiple local maxima for normal mixture models. In this case of unrestricted component covariance matrices, $L(\Psi)$ is unbounded, as each data point gives rise to a singularity on the edge of the parameter space; see, for example, McLachlan and Basford (1988, Chapter 2), McLachlan and Peel (2000a, Section 3.8), Gelman, Carlin, Stern, and Rubin (2004), and Lindsay (1995). It suffices to direct attention to local maxima in the interior of the parameter space, as under essentially the usual regularity conditions (Kiefer, 1978; Peters and Walker, 1978), there exists a sequence of roots of the likelihood equation that is consistent and asymptotically efficient. With probability tending to one, these roots correspond to local maxima in the interior of the parameter space. In practice, however, consideration has to be given to the problem of relatively large local maxima that occur as a consequence of a fitted component having a very small (but nonzero) variance for univariate data or generalized variance (the determinant of the covariance matrix) for multivariate data. Such a component corresponds to a cluster containing a few data points either relatively close together or almost lying in a lower dimensional subspace in the case of multivariate data. There is thus a need to monitor the relative size of the component variances for univariate observations and of the generalized component variances for multivariate data in an attempt to identify these spurious local maximizers. Hathaway (1983, 1985, 1986) considers a constrained formulation of this problem in order to avoid singularities and to reduce the number of spurious local maximizers. Of course the possibility here that for a given starting point the EM algorithm may converge to a spurious local maximizer or may not converge at all is not a failing of this algorithm. Rather it is a consequence of the properties of the likelihood function for the normal mixture model with unrestricted component matrices in the case of ungrouped data. Section 3.5.4 discusses this issue a little further.

2.7.3 Numerical Example: Red Blood Cell Volume Data

We consider here a data set from McLachlan and Jones (1988) in their study of the change in the red blood cell populations of cows exposed to the tick-borne parasite *Anaplasma marginale* in a laboratory trial. The data collected were in the form of red blood cell volume distributions obtained from a Coulter counter. The observed cell counts so obtained from a cow 21 days after inoculation are listed in Table 2.11, along with the expected cell frequencies as obtained by McLachlan and Jones (1988) for this data set under the normal mixture model below. The lower and upper truncation values for these red blood cell volume counts are 28.8 fl and 158.4 fl, respectively, and the grouping is into 18 intervals of equal width of 7.2 fl. A cursory inspection of the observed cell counts displayed in histogram form on the logarithmic scale in Figure 2.3 suggests that the red blood cell volume is bimodal at 21 days after inoculation. Accordingly, McLachlan and Jones (1988) modeled the log of the red blood cell volume distribution by a two-component normal density,

$$f(w; \Psi) = \pi_1 \phi(w; \mu_1, \sigma_1^2) + \pi_2 \phi(w; \mu_2, \sigma_2^2). \quad (2.59)$$

McLachlan and Jones (1988) fitted this model to the data in their original grouped and truncated form, and their results are to be reported in Section 2.8.7, which is devoted to ML estimation from data grouped into intervals. But to illustrate the fitting of a normal mixture model to individual data points, we fit this model by taking the observations within an interval to be at its midpoint and also ignoring the extreme intervals. The adequacy of this approximation depends on the ratio of the width of the intervals relative to the variance; see Heitjan (1989).

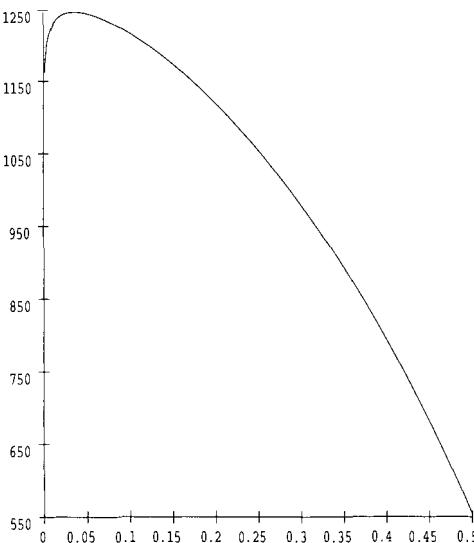


Figure 2.3 Plot of histogram of red blood cell volume data.

From the initial value of $\Psi^{(0)}$ as specified in Table 2.12, the EM algorithm converged after 65 iterations. The stopping criterion was based on the relative change in the log likelihood, using a tolerance value of 10^{-8} .

2.8 EXAMPLE 2.9: GROUPED AND TRUNCATED DATA

2.8.1 Introduction

We consider now the application of the EM algorithm to continuous data that are grouped into intervals and may also be truncated. The results are then specialized to the case of univariate normal data. In general with ML estimation for grouped data, Dempster and Rubin (1983) show how Shepherd's corrections arise through application of the EM algorithm; see Heitjan (1989). Specific examples of ML estimation for grouped discrete data may be found, for instance, in Schader and Schmid (1985) and Adamidis and Loukas (1993), who consider the fitting of the negative binomial distribution and Poisson mixtures in binomial proportions, respectively. More recently, Cadez, Smyth, McLachlan, and McLaren (2002) considered ML estimation for grouped and truncated multivariate data.

2.8.2 Specification of Complete Data

Let W be a random variable with p.d.f. $f(w; \Psi)$ specified up to a vector Ψ of unknown parameters. Suppose that the sample space \mathcal{W} of W is partitioned into v mutually exclusive intervals \mathcal{W}_j ($j = 1, \dots, v$). Independent observations are made on W , but only the

Table 2.11 Doubly Truncated Red Blood Cell Volume Data for a Cow 21 Days after Inoculation.

Group no.	Lower endpoint of cell volume interval	Frequencies	
		Observed	Expected
1	28.8	10	6.5
2	36.0	21	27.0
3	43.2	51	54.8
4	50.4	77	70.0
5	57.6	70	66.4
6	64.8	50	53.9
7	72.0	44	44.2
8	79.2	40	42.1
9	86.4	46	45.9
10	93.6	54	50.8
11	100.8	53	53.2
12	108.0	54	51.4
13	115.2	44	45.8
14	122.4	36	38.2
15	129.6	29	29.9
16	136.8	21	22.3
17	144.0	16	15.9
18	151.2	13	10.9

Source: Adapted from McLachlan and Jones (1988), with permission of the Biometric Society.

Table 2.12 Results of EM Algorithm Applied to Midpoints of the Class Intervals for Red Blood Cell Volume Data.

Parameter	Initial Value	MLE
π_1	0.45	0.5192
μ_1	4.00	4.1103
μ_2	4.40	4.7230
σ_1^2	0.08	0.0685
σ_2^2	0.05	0.0286

number n_j falling in \mathcal{W}_j ($j = 1, \dots, r$) is recorded, where $r \leq v$. That is, individual observations are not recorded but only the class intervals \mathcal{W}_j in which they fall are recorded; further, even such observations are made only if the W value falls in one of the intervals \mathcal{W}_j ($j = 1, \dots, r$).

For given

$$n = \sum_{j=1}^r n_j,$$

the observed data

$$\mathbf{y} = (n_1, \dots, n_r)^T$$

has a multinomial distribution, consisting of n draws on r categories with probabilities $P_j(\Psi)/P(\Psi)$, $j = 1, \dots, r$, where

$$P_j(\Psi) = \int_{\mathcal{W}_j} f(w; \Psi) dw \quad (2.60)$$

and

$$P(\Psi) = \sum_{j=1}^r P_j(\Psi).$$

Thus the (incomplete-data) log likelihood is given by

$$\log L(\Psi) = \sum_{j=1}^r n_j \log\{P_j(\Psi)/P(\Psi)\} + c_1, \quad (2.61)$$

where

$$c_1 = \log\{n!/\prod_{j=1}^r n_j!\}.$$

We can solve this problem within the EM framework by introducing the vectors

$$\mathbf{u} = (n_{r+1}, \dots, n_v)^T \quad (2.62)$$

and

$$\mathbf{w}_j = (w_{j1}, \dots, w_{jn_j})^T \quad (j = 1, \dots, v) \quad (2.63)$$

as the missing data. The vector \mathbf{u} contains the unobservable frequencies in the case of truncation ($r < v$), while \mathbf{w}_j contains the n_j unobservable individual observations in the j th interval \mathcal{W}_j ($j = 1, \dots, v$).

The complete-data vector \mathbf{x} corresponding to the missing data defined by (2.62) and (2.63) is

$$\mathbf{x} = (\mathbf{y}^T, \mathbf{u}^T, \mathbf{w}_1^T, \dots, \mathbf{w}_v^T)^T.$$

These missing data have been introduced so that the complete-data log likelihood function for Ψ , $\log L_c(\Psi)$, is equivalent to

$$\sum_{j=1}^v \sum_{l=1}^{n_j} \log f(w_{jl}; \Psi), \quad (2.64)$$

which is the log likelihood function formed under the assumption that the observations in $\mathbf{w}_1, \dots, \mathbf{w}_v$ constitute an observed random sample of size $n + m$ from $f(w; \Psi)$ on \mathcal{W} , where

$$m = \sum_{j=r+1}^v n_j.$$

We now consider the specification of the distributions of the missing data \mathbf{u} and $\mathbf{w}_1, \dots, \mathbf{w}_v$, so that $\log L_c(\Psi)$ is equivalent to (2.64) and hence implies the incomplete-data log likelihood, $\log L(\Psi)$, as given by (2.61). We shall proceed by considering $L_c(\Psi)$ as the product of the density of \mathbf{y} , the conditional density of \mathbf{u} given \mathbf{y} , and the conditional density of $\mathbf{w}_1, \dots, \mathbf{w}_v$ given \mathbf{y} and \mathbf{u} . Concerning the latter, $\mathbf{w}_1, \dots, \mathbf{w}_v$ are taken to be independent with the n_j observations w_{jk} ($k = 1, \dots, n_j$) in \mathbf{w}_j , constituting an observed random sample of size n_j from the density

$$h_j(w; \Psi) = f(w; \Psi)/P_j(\Psi) \quad (j = 1, \dots, v). \quad (2.65)$$

It remains now to specify the conditional distribution of \mathbf{u} given \mathbf{y} , which we shall write as $d(\mathbf{u} \mid \mathbf{y}; \Psi)$. This is one instance with the application of the EM algorithm where the specification of the distribution of the missing data is perhaps not straightforward. Hence we shall leave it unspecified for the moment and proceed with the formation of $L_c(\Psi)$ in the manner described above. It follows that

$$\log L_c(\Psi) = \log L(\Psi) + \log d(\mathbf{u} \mid \mathbf{y}; \Psi) + \sum_{j=1}^v \sum_{l=1}^{n_j} \log h_j(w_{jl}; \Psi),$$

which, on using (2.61) and (2.65), equals

$$\begin{aligned} \log L_c(\Psi) &= \sum_{j=1}^r n_j \log \{P_j(\Psi)/P(\Psi)\} + c_1 \\ &\quad + \log d(\mathbf{u} \mid \mathbf{y}; \Psi) + \sum_{j=1}^v \sum_{l=1}^{n_j} \log \{f(w_{jl}; \Psi)/P_j(\Psi)\} \\ &= \sum_{j=1}^v \sum_{l=1}^{n_j} \log f(w_{jl}; \Psi) \\ &\quad + \log d(\mathbf{u} \mid \mathbf{y}; \Psi) - \log \{P(\Psi)\}^n \prod_{j=r+1}^v P_j\{\{(\Psi)\}^{n_j}\} \\ &\quad + c_1. \end{aligned} \tag{2.66}$$

Thus (2.66) is equivalent to (2.64), and so implies the incomplete-data log likelihood (2.61), if the conditional distribution of \mathbf{u} given \mathbf{y} is specified to be

$$d(\mathbf{u} \mid \mathbf{y}; \Psi) = c_2 \{P(\Psi)\}^n \prod_{j=r+1}^v \{P_j(\Psi)\}^{n_j}, \tag{2.67}$$

where c_2 is a normalizing constant that does not depend on Ψ . It can be shown that the right-hand side of (2.67) defines a proper probability function if

$$c_2 = (m + n - 1)! / \{(n - 1)! \prod_{j=r+1}^v n_j!\}.$$

The use of (2.67) can be viewed as simply a device to produce the desired form (2.64) for the complete-data likelihood $L_c(\Psi)$. The distribution (2.67) reduces to the negative binomial in the case $r = v - 1$; see Achuthan and Krishnan (1992) for an example of its use in a genetics problem with truncation.

2.8.3 E-Step

On working with (2.64) as the complete-data log likelihood function for Ψ , the E-step on the $(k + 1)$ th iteration is effected by first taking the expectation of $\log L_c(\Psi)$ conditional also on w_1, \dots, w_v and \mathbf{u} , as well as \mathbf{y} . On taking the expectation of $\log L_c(\Psi)$ over w_1, \dots, w_v and finally \mathbf{u} , it follows that on the $(k + 1)$ th iteration, the expectation of (2.64) conditional on \mathbf{y} is given by

$$Q(\Psi; \Psi^{(k)}) = \sum_{j=1}^v n_j^{(k)} Q_j(\Psi; \Psi^{(k)}), \tag{2.68}$$

where

$$Q_j(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log f(W; \Psi) \mid W \in \mathcal{W}_j \} \quad (2.69)$$

and where

$$\begin{aligned} n_j^{(k)} &= n_j \quad (j = 1, \dots, r) \\ &= E_{\Psi^{(k)}}(n_j \mid \mathbf{y}) \quad (j = r + 1, \dots, v). \end{aligned}$$

With the conditional distribution of \mathbf{u} given \mathbf{y} specified by (2.67), it can be shown that

$$E_{\Psi^{(k)}}(n_j \mid \mathbf{y}) = n P_j(\Psi^{(k)}) / P(\Psi^{(k)}) \quad (j = r + 1, \dots, v). \quad (2.70)$$

We have seen that a crude approach to estimation from grouped data is to form the likelihood function with the n_j observations in the j th interval \mathcal{W}_j taken to be at the midpoint

$$\bar{w}_j = \frac{1}{2}(w_{j-1} + w_j)$$

with (or without) an appropriate adjustment to handle truncation or the extreme intervals in the case of infinite endpoints. With this approach the j th interval contributes

$$n_j \log f(\bar{w}; \Psi)$$

to the log likelihood. We see from (2.68) that with the EM algorithm the corresponding contribution to the log likelihood is

$$n_j E_{\Psi^{(k)}} \{ \log f(W; \Psi) \mid W \in \mathcal{W}_j \}.$$

Thus the EM algorithm uses the current conditional expectation of the log density over the j th interval. The midpoint approach simply approximates this expectation by evaluating the log density at the midpoint of the interval.

2.8.4 M-Step

On the M-step at the $(k+1)$ th iteration, we have on differentiation of $Q(\Psi; \Psi^{(k)})$ with respect to Ψ that $\Psi^{(k+1)}$ is a root of the equation

$$\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi = \sum_{j=1}^v n_j^{(k)} \partial Q_j(\Psi; \Psi^{(k)}) / \partial \Psi, \quad (2.71)$$

where

$$\partial Q_j(\Psi; \Psi^{(k)}) / \partial \Psi = E_{\Psi^{(k)}} \{ \partial \log f(W; \Psi) / \partial \Psi \mid W \in \mathcal{W}_j \}, \quad (2.72)$$

on interchanging the operations of differentiation and expectation.

2.8.5 Confirmation of Incomplete-Data Score Statistic

It is of interest in this example to check that the expression

$$[\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi^{(k)} = \Psi}$$

for the incomplete-data score statistic $S(\mathbf{y}; \Psi)$ in terms of the complete-data specification agrees with what would be obtained by direct differentiation of the incomplete-data log likelihood $\log L(\Psi)$ as given by (2.61).

Now on evaluating the right-hand side of (2.72) at $\Psi^{(k)} = \Psi$, it follows that

$$\begin{aligned} [\partial Q_j(\Psi; \Psi^{(k)})/\partial \Psi]_{\Psi^{(k)}=\Psi} &= \{\partial P_j(\Psi)/\partial \Psi\}/P_j(\Psi) \\ &= \partial \log P_j(\Psi)/\partial \Psi. \end{aligned} \quad (2.73)$$

Considering the summation in (2.71) over the truncated intervals, we have from (2.70) and (2.73) that

$$\begin{aligned} &\left[\sum_{j=r+1}^v n_j^{(k)} \partial Q_j(\Psi; \Psi^{(k)})/\partial \Psi \right]_{\Psi^{(k)}=\Psi} \\ &= \{n/P(\Psi)\} \sum_{j=r+1}^v \partial P_j(\Psi)/\partial \Psi \end{aligned} \quad (2.74)$$

$$= -\{n/P(\Psi)\} \partial P(\Psi)/\partial \Psi \quad (2.75)$$

$$= -n \partial \log P(\Psi)/\partial \Psi. \quad (2.76)$$

The result (2.75) follows from (2.74) since

$$\sum_{j=r+1}^v P_j(\Psi) = 1 - P(\Psi).$$

On using (2.73) and (2.76) in (2.71), we have that

$$\begin{aligned} S(\mathbf{y}; \Psi) &= [\partial Q(\Psi; \Psi^{(k)})/\partial \Psi]_{\Psi^{(k)}=\Psi} \\ &= \sum_{j=1}^r n_j \partial \log P_j(\Psi)/\partial \Psi \\ &\quad - n \partial \log P(\Psi)/\partial \Psi, \end{aligned} \quad (2.77)$$

which obviously agrees with the expression for $S(\mathbf{y}; \Psi)$ obtained by differentiating $\log L(\Psi)$ directly with respect to Ψ .

We can express $\partial \log P(\Psi)/\partial \Psi$ in (2.77) as

$$n \partial \log P(\Psi)/\partial \Psi = \sum_{j=1}^r \{nP_j(\Psi)/P(\Psi)\} \partial \log P_j(\Psi)/\partial \Psi,$$

and so $S(\mathbf{y}; \Psi)$ can be expressed in the form

$$S(\mathbf{y}; \Psi) = \sum_{j=1}^r \left\{ n_j - \frac{n P_j(\Psi)}{P(\Psi)} \right\} \partial \log P_j(\Psi)/\partial \Psi. \quad (2.78)$$

We shall make use of this result in Section 4.4.

2.8.6 M-Step for Grouped Normal Data

To discuss the M-step, we consider the case where

$$f(w; \Psi) = \phi(w; \mu, \sigma^2)$$

denotes the normal density with mean μ and variance σ^2 and

$$\Psi = (\mu, \sigma^2)^T.$$

In this case,

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) &= -\frac{1}{2}(n + \sum_{j=r+1}^v n_j^{(k)})\{\log(2\pi) + \log \sigma^2\} \\ &\quad -\frac{1}{2}\sigma^{-2} \sum_{j=1}^v n_j^{(k)} E_{\Psi^{(k)}}\{(W - \mu)^2 \mid W \in \mathcal{W}_j\}. \end{aligned}$$

It can now be shown that the value $\Psi^{(k+1)}$ of Ψ that maximizes $Q(\Psi, \Psi^{(k)})$ is given by

$$\Psi^{(k+1)} = (\mu^{(k+1)}, \sigma^{(k+1)^2})^T,$$

where

$$\mu^{(k+1)} = \sum_{j=1}^v n_j^{(k)} \{E_{\Psi^{(k)}}(W \mid W \in \mathcal{W}_j)\} / \sum_{j=1}^v n_j^{(k)} \quad (2.79)$$

and

$$\sigma^{(k+1)^2} = \sum_{j=1}^v n_j^{(k)} [E_{\Psi^{(k)}}\{(W - \mu^{(k+1)})^2 \mid W \in \mathcal{W}_j\}] / \sum_{j=1}^v n_j^{(k)}. \quad (2.80)$$

We see from (2.79) and (2.80) that in computing $\mu^{(k+1)}$ and $\sigma^{(k+1)^2}$, an unobservable frequency n_j in the case of truncation is replaced by

$$n_j^{(k)} = n P_j(\Psi^{(k)}) / P(\Psi^{(k)}).$$

For the computation of the mean $\mu^{(k+1)}$, the individual values of W in the interval \mathcal{W}_j are replaced by the current conditional expectation of W given W falls in \mathcal{W}_j . Similarly, for the computation of the variance $\sigma^{(k+1)^2}$, the individual values of $(W - \mu^{(k+1)})^2$ in the interval \mathcal{W}_j are replaced by the current conditional expectation of $(W - \mu^{(k+1)})^2$ given W falls in \mathcal{W}_j . Jones and McLachlan (1990) have developed an algorithm for fitting finite mixtures of normal distributions to data that are grouped into intervals and that may also be truncated.

2.8.7 Numerical Example: Grouped Log Normal Data

In the previous section, we fitted on the log scale a two-component normal mixture model (2.59) to the midpoints of the class intervals in Table 2.8 into which the red blood cell volume have been grouped. We now report the results of McLachlan and Jones (1988) who fitted this model to the data in their original grouped and truncated form. In Section 2.7, we described the fitting of a normal mixture density, while in Section 2.8 we described the fitting of a single normal density to grouped and truncated data. These two M-steps can be combined to give the M-step for the fitting of a normal mixture density to data that are grouped and truncated, as detailed in McLachlan and Jones (1988). In Table 2.13, we display the results of the fit that they obtained for the present data set after 117 iterations of the EM algorithm started from the same value for Ψ as in Table 2.12. Also displayed there is the corresponding fit obtained in the previous section for the midpoint approximation

method. The entries in parentheses refer to the standard errors of the above estimates and they were obtained using methodology to be discussed in Chapter 4. The standard errors for the estimates from the midpoint approximation method were obtained by using the inverse of the empirical information matrix (4.41) to approximate the covariance matrix of the MLE, while for the exact MLE method, the corresponding form (4.45) of the empirical information matrix for grouped data was used.

Table 2.13 Results of EM Algorithm Applied to Red Blood Cell Volume Data in Grouped and Truncated Form on the Log Scale.

Parameter	MLE	
	Midpoint Approximation	Exact Result
π_1	0.5192 (0.0417)	0.4521 (0.0521)
μ_1	4.1103 (0.0326)	4.0728 (0.0384)
μ_2	4.7230 (0.0175)	4.7165 (0.0242)
σ_1^2	0.0685 (0.0105)	0.0575 (0.0107)
σ_2^2	0.0286 (0.0047)	0.0438 (0.0099)

It can be seen from Table 2.13 that the midpoint approximation technique overestimates the mixing parameter. The estimates of the two component means are similar for the midpoint approximation and exact ML methods, while the former overestimates the first component variance and underestimates the second component variance. The approximate standard errors of the parameter estimates are lower for the midpoint approximation method.

The expected frequencies for this model fit are displayed alongside the observed frequencies in Table 2.11. The chi-squared goodness-of-fit statistic is equal to 5.77 on twelve degrees of freedom, confirming that the two-component log normal mixture provides an adequate fit to the observed frequency counts.

2.9 EXAMPLE 2.10: A HIDDEN MARKOV AR(1) MODEL

Examples 2.7 and 2.8 dealt with mixtures of normal distributions when the observations are independently generated from the same mixture. Following Hamilton (1989, 1990, 1993, 1994), we consider an example of a mixture of two AR(1) series, the parameters of the mixture being governed by a hidden (that is, unobservable) two-state Markovian regime. Consider a time series $\{w_j, j = 1, 2, \dots\}$ of the form

$$w_j - \mu_{s_j} = \beta_1(w_{j-1} - \mu_{s_{j-1}}) + \epsilon_j, \quad (2.81)$$

where $s_j = i$ if w_j is in state S_i ($i = 1, 2$) of the Markov chain, and where the $\epsilon_j \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. We let $z_{ij} = 1$ if $s_j = i$ ($i = 1, 2; j = 1, \dots, n$), and put

$$\mathbf{z} = (z_1^T, \dots, z_n^T)^T,$$

where $(z_j)_i = z_{ij}$. The dependence between the unobservable (hidden) state-indicators z_j associated with the w_j is specified by a stationary Markov chain with transition probability matrix $\mathbf{A} = ((a_{hi}))$, $h, i = 1, 2$, where

$$a_{hi} = \text{pr}\{Z_{i,j+1} = 1 \mid Z_{hj} = 1\}. \quad (2.82)$$

The initial distribution of the Markov chain is specified by π_{i1} ($i = 1, 2$). We let

$$\mathbf{y}_j = (w_1, w_2, \dots, w_j)^T,$$

and suppose that the observed data vector \mathbf{y} is equal to \mathbf{y}_n .

Here it is assumed that $\mu_1 = \mu_2 = 0$ and that the state-conditional distribution of w_j given \mathbf{y}_{j-1} is given by

$$w_j \mid \mathbf{y}_{j-1}, s_j \sim N(\beta_{s_j} w_{j-1}, \sigma^2). \quad (2.83)$$

We put $\boldsymbol{\theta} = (\beta_1, \beta_2, \sigma^2)^T$, so that the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the elements of \mathbf{A} in addition to $\boldsymbol{\theta}$. Also, we let $f_i(w_j \mid \mathbf{y}_{j-1}; \boldsymbol{\theta})$ denote the conditional density of w_j given \mathbf{y}_{j-1} and $s_j = i$. It is specified by (2.83).

The EM algorithm can be applied to this problem, treating the vector \mathbf{z} of hidden state-indicators as the missing data. Analogous to the formulation of the EM framework in the normal mixture problem, the complete-data log likelihood is given by

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^2 \sum_{j=1}^n z_{ij} \log f_i(w_j \mid \mathbf{y}_{j-1}; \boldsymbol{\theta}) + \text{pr}_{\boldsymbol{\Psi}}(\mathbf{z}). \quad (2.84)$$

On taking the conditional expectation of the complete-data log likelihood $\log L_c(\boldsymbol{\Psi})$ given by (2.84), using $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$, we have that

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^2 \sum_{j=1}^n \tau_{ij}^{(k)} \log f_i(w_j \mid \mathbf{y}_{j-1}; \boldsymbol{\Psi}) + \text{pr}_{\boldsymbol{\Psi}^{(k)}}(\mathbf{z}), \quad (2.85)$$

where $\tau_{ij}^{(k)} = \tau_{ij}(\mathbf{y}_n; \boldsymbol{\Psi}^{(k)})$ and

$$\begin{aligned} \tau_{ij}(\mathbf{y}_n; \boldsymbol{\Psi}) &= E_{\boldsymbol{\Psi}}(Z_{ij} \mid \mathbf{y}_n) \\ &= \text{pr}_{\boldsymbol{\Psi}}\{Z_{ij} = 1 \mid \mathbf{y}_n\}. \end{aligned} \quad (2.86)$$

In order to carry out the E-step, we essentially have to be able to compute the probability

$$\xi_j(h, i) = \text{pr}_{\boldsymbol{\Psi}}\{Z_{hj} = 1, Z_{i,j+1} = 1 \mid \mathbf{y}_n\} \quad (2.87)$$

for $h, i = 1, 2$ and $j = 1, \dots, n - 1$; see also Section 8.2, where the use of the EM algorithm in a hidden Markov chain is considered further.

Concerning the M-step, the updated estimate of the autoregressive parameter β_i is a solution of

$$\sum_{j=1}^n \tau_{ij}^{(k)} (w_j - \hat{\beta}_i y_{j-1}) w_{j-1} \quad (i = 1, 2). \quad (2.88)$$

This gives

$$\beta_i^{(k+1)} = \sum_{j=1}^n w_j^{(k)} \tilde{w}_j^{(k)} / \sum_{j=1}^n w_j^{(k)2} \quad (i = 1, 2), \quad (2.89)$$

where

$$w_j^{(k)} = w_j \sqrt{\tau_{ij}^{(k)}}$$

and

$$\tilde{w}_j^{(k)} = w_{j-1} \sqrt{\tau_{ij}^{(k)}}.$$

The updated estimate of σ^2 is given by

$$\sigma^{(k+1)^2} = \frac{1}{n} \sum_{i=1}^2 \sum_{j=1}^n \tau_{ij}^{(k)} (w_j - \beta_i^{(k+1)} w_{j-1})^2. \quad (2.90)$$

Hamilton (1990) has shown that if $\hat{\Psi}$ denotes the MLE of Ψ , then the MLEs of a_{hi} satisfy the equation

$$a_{hi} = \frac{\sum_{j=1}^{n-1} \text{pr}_{\hat{\Psi}}\{Z_{hj} = 1, Z_{i,j+1} = 1 \mid \mathbf{y}_n\}}{\sum_{j=1}^{n-1} \text{pr}_{\hat{\Psi}}\{Z_{hj} = 1 \mid \mathbf{y}_n\}}. \quad (2.91)$$

Thus the estimates of the transition probabilities a_{hi} can be updated using (2.91). It corresponds to equation (8.4) in Section 8.2 on hidden Markov models.

For other applications of the EM algorithm to time series analysis, see Brockwell and Davis (1996) (missing observations in an AR(2) model) and Shumway and Stoffer (1982, 2000) (AR(1) process with observational noise, where the E-step is Kalman filtering for smoothing, and M-step is the usual multivariate normal regression estimation). Other econometric applications of the EM algorithm may be found in Kiefer (1980), Watson and Engle (1983), and Ruud (1991).

This Page Intentionally Left Blank

CHAPTER 3

BASIC THEORY OF THE EM ALGORITHM

3.1 INTRODUCTION

Having illustrated the EM algorithm in a variety of problems and having presented a general formulation of it in earlier chapters, we now begin a systematic account of its theory. In this chapter, we show that the likelihood increases with each EM or GEM iteration, and, under fairly general conditions, the likelihood values converge to stationary values. However, in order to caution the reader that the EM algorithm will not always lead to even a local if not a global maximum of the likelihood, we present two examples in one of which convergence is to a saddle point and in the other to a local *minimum* of the likelihood. The principles of Self-Consistency and Missing Information are explained in the sequel. The notion that the rate at which the EM algorithm converges depends upon the amount of information missing in the incomplete data *vis a vis* the formulated complete data, is made explicit by deriving results regarding the rate of convergence in terms of information matrices for the incomplete- and complete-data problems. Theoretical results are illustrated mostly with examples discussed in earlier chapters.

Our intent here is to collect in one place and present in a unified manner the available results concerning the basic theory of EM—aspects of convergence and rates of convergence. The treatment in this chapter is inevitably mathematical. Readers not interested in mathematical details may skip proofs of theorems.

3.2 MONOTONICITY OF THE EM ALGORITHM

Dempster, Laird, and Rubin (1977) showed that the (incomplete-data) likelihood function $L(\Psi)$ is not decreased after an EM iteration; that is,

$$L(\Psi^{(k+1)}) \geq L(\Psi^{(k)}) \quad (3.1)$$

for $k = 0, 1, 2, \dots$. To see this, let

$$k(\mathbf{x} | \mathbf{y}; \Psi) = g_c(\mathbf{x}; \Psi)/g(\mathbf{y}; \Psi) \quad (3.2)$$

be the conditional density of \mathbf{X} given $\mathbf{Y} = \mathbf{y}$. Then the log likelihood is given by

$$\begin{aligned} \log L(\Psi) &= \log g(\mathbf{y}; \Psi) \\ &= \log g_c(\mathbf{x}; \Psi) - \log k(\mathbf{x} | \mathbf{y}; \Psi) \\ &= \log L_c(\Psi) - \log k(\mathbf{x} | \mathbf{y}; \Psi). \end{aligned} \quad (3.3)$$

On taking the expectations of both sides of (3.3) with respect to the conditional distribution of \mathbf{X} given $\mathbf{Y} = \mathbf{y}$, using the fit $\Psi^{(k)}$ for Ψ , we have that

$$\begin{aligned} \log L(\Psi) &= E_{\Psi^{(k)}} \{ \log L_c(\Psi) | \mathbf{y} \} \\ &\quad - E_{\Psi^{(k)}} \{ \log k(\mathbf{X} | \mathbf{y}; \Psi) | \mathbf{y} \} \\ &= Q(\Psi; \Psi^{(k)}) - H(\Psi; \Psi^{(k)}), \end{aligned} \quad (3.4)$$

where

$$H(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log k(\mathbf{X} | \mathbf{y}; \Psi) | \mathbf{y} \}. \quad (3.5)$$

From (3.4), we have that

$$\begin{aligned} &\log L(\Psi^{(k+1)}) - \log L(\Psi^{(k)}) \\ &= \{Q(\Psi^{(k+1)}; \Psi^{(k)}) - Q(\Psi^{(k)}; \Psi^{(k)})\} \\ &\quad - \{H(\Psi^{(k+1)}; \Psi^{(k)}) - H(\Psi^{(k)}; \Psi^{(k)})\}. \end{aligned} \quad (3.6)$$

The first difference on the right-hand side of (3.6) is nonnegative since $\Psi^{(k+1)}$ is chosen so that

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) \geq Q(\Psi; \Psi^{(k)}) \quad (3.7)$$

for all $\Psi \in \Omega$. Hence (3.1) holds if the second difference on the right-hand side of (3.6) is nonpositive; that is, if

$$H(\Psi^{(k+1)}; \Psi^{(k)}) - H(\Psi^{(k)}; \Psi^{(k)}) \leq 0. \quad (3.8)$$

Now for any Ψ ,

$$\begin{aligned} &H(\Psi; \Psi^{(k)}) - H(\Psi^{(k)}; \Psi^{(k)}) \\ &= E_{\Psi^{(k)}} [\log \{k(\mathbf{X} | \mathbf{y}; \Psi)/k(\mathbf{X} | \mathbf{y}; \Psi^{(k)})\} | \mathbf{y}] \\ &\leq \log [E_{\Psi^{(k)}} \{k(\mathbf{X} | \mathbf{y}; \Psi)/k(\mathbf{X} | \mathbf{y}; \Psi^{(k)})\} | \mathbf{y}] \end{aligned} \quad (3.9)$$

$$\begin{aligned} &= \log \int_{\mathcal{X}(\mathbf{y})} k(\mathbf{x} | \mathbf{y}; \Psi) d\mathbf{x} \\ &= 0, \end{aligned} \quad (3.10)$$

where the inequality in (3.9) is a consequence of Jensen's inequality and the concavity of the logarithmic function.

This establishes (3.8), and hence the inequality (3.1), showing that the likelihood $L(\Psi)$ is not decreased after an EM iteration. The likelihood will be increased if the inequality (3.7) is strict. Thus for a bounded sequence of likelihood values $\{L(\Psi^{(k)})\}$, $L(\Psi^{(k)})$ converges monotonically to some L^* .

A consequence of (3.1) is the self-consistency of the EM algorithm. For if the MLE $\hat{\Psi}$ globally maximizes $L(\Psi)$, it must satisfy

$$Q(\hat{\Psi}; \hat{\Psi}) \geq Q(\Psi; \hat{\Psi}) \quad (3.11)$$

for all Ψ . Otherwise

$$Q(\hat{\Psi}; \hat{\Psi}) < Q(\Psi_o; \hat{\Psi})$$

for some Ψ_o , implying that

$$L(\Psi_o) > L(\hat{\Psi}),$$

which would contradict the fact that $\hat{\Psi}$ is the global maximizer of $L(\Psi)$.

It will be seen that the differential form of (3.11) is that $\hat{\Psi}$ is a root of the equation

$$[\partial Q(\Psi; \hat{\Psi}) / \partial \Psi]_{\Psi=\hat{\Psi}} = \mathbf{0}.$$

This equation will be established shortly in the next section. As noted by Efron (1982), the self-consistency property of the MLE has been rediscovered in many different contexts since Fisher's (1922, 1925, 1934) original papers. Here it forms the basis of the EM algorithm.

3.3 MONOTONICITY OF A GENERALIZED EM ALGORITHM

From the definition of a GEM algorithm given in Section 1.5.5, $\Psi^{(k+1)}$ is chosen not to globally maximize $Q(\Psi; \Psi^{(k)})$ with respect to Ψ , but rather to satisfy

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) \geq Q(\Psi^{(k)}; \Psi^{(k)}). \quad (3.12)$$

We have seen in the previous section that this condition (3.12) is sufficient to ensure that (3.1) holds for an iterative sequence $\{\Psi^{(k)}\}$. Thus the likelihood is not decreased after a GEM iteration.

3.4 CONVERGENCE OF AN EM SEQUENCE TO A STATIONARY VALUE

3.4.1 Introduction

As shown in the last section, for a sequence of likelihood values $\{L(\Psi^{(k)})\}$ bounded above, $L(\Psi^{(k)})$ converges monotonically to some value L^* . In almost all applications, L^* is a stationary value. That is, $L^* = L(\Psi^*)$ for some point Ψ^* at which

$$\partial L(\Psi) / \partial \Psi = \mathbf{0},$$

or equivalently,

$$\partial \log L(\Psi) / \partial \Psi = \mathbf{0}.$$

Moreover, in many practical applications, L^* will be a local maximum. In any event, if an EM sequence $\{\Psi^{(k)}\}$ is trapped at some stationary point Ψ^* that is not a local or global maximizer of $L(\Psi)$ (for example, a saddle point), a small random perturbation of Ψ away from the saddle point Ψ^* will cause the EM algorithm to diverge from the saddle point.

In general, if $L(\Psi)$ has several stationary points, convergence of the EM sequence to either type (local or global maximizers, saddle points) depends on the choice of starting point $\Psi^{(0)}$. In the case where the likelihood function $L(\Psi)$ is unimodal in Ω (and a certain differentiability condition is satisfied), any EM sequence converges to the unique MLE, irrespective of its starting point $\Psi^{(0)}$. This result is to be given at the end of the next section as a corollary to the convergence theorems to be presented in the sequel.

On differentiating both sides of (3.4), we have that

$$\partial \log L(\Psi)/\partial \Psi = \partial Q(\Psi; \Psi^{(k)})/\partial \Psi - \partial H(\Psi; \Psi^{(k)})/\partial \Psi. \quad (3.13)$$

The inequality (3.10) implies that

$$H(\Psi; \Psi^{(k)}) \leq H(\Psi^{(k)}; \Psi^{(k)})$$

for all $\Psi \in \Omega$, and so

$$[\partial H(\Psi; \Psi^{(k)})/\partial \Psi]_{\Psi=\Psi^{(k)}} = 0. \quad (3.14)$$

Let Ψ_o be some arbitrary value of Ψ . On putting $\Psi^{(k)} = \Psi_o$ in (3.13), we have from (3.14) that

$$\partial \log L(\Psi_o)/\partial \Psi = [\partial Q(\Psi; \Psi_o)/\partial \Psi]_{\Psi=\Psi_o}, \quad (3.15)$$

where $\partial \log L(\Psi_o)/\partial \Psi$ denotes $\partial \log L(\Psi)/\partial \Psi$ evaluated at $\Psi = \Psi_o$.

Suppose that $\Psi = \Psi^*$, where Ψ^* is a stationary point of $L(\Psi)$. Then from (3.15),

$$\begin{aligned} \partial \log L(\Psi^*)/\partial \Psi &= [\partial Q(\Psi; \Psi^*)/\partial \Psi]_{\Psi=\Psi^*} \\ &= 0. \end{aligned} \quad (3.16)$$

It can be seen from (3.16) that the EM algorithm can converge to a saddle point Ψ^* if $Q(\Psi; \Psi^*)$ is globally maximized over $\Psi \in \Omega$ at Ψ^* .

This led Wu (1983) to propose the condition

$$\sup_{\Psi \in \Omega} Q(\Psi; \Psi^*) > Q(\Psi^*; \Psi^*) \quad (3.17)$$

for any stationary point Ψ^* that is not a local maximizer of $L(\Psi)$. This condition in conjunction with his regularity conditions to be given in the next subsection will ensure that all limit points of any instance of the EM algorithm are local maximizers of $L(\Psi)$ and that $L(\Psi^{(k)})$ converges monotonically to $L^* = L(\Psi^*)$ for some local maximizer Ψ^* . However, the utility of condition (3.17) is limited given that it is typically hard to verify.

In Section 3.5, we shall give an example of an EM sequence converging to a saddle point, and also an example of an EM sequence that converges to a local minimizer if started from some isolated initial values. A reader not concerned with these esoteric issues of the EM algorithm may wish to proceed directly to Section 3.7.

3.4.2 Regularity Conditions of Wu (1983)

As can be seen from (1.54), the M-step of the EM algorithm involves the point-to-set map

$$\mathcal{M}(\Psi^{(k)}) = \arg \max_{\Psi \in \Omega} Q(\Psi; \Psi^{(k)});$$

that is, $\mathcal{M}(\Psi^{(k)})$ is the set of values of Ψ that maximize $Q(\Psi; \Psi^{(k)})$ over $\Psi \in \Omega$. For a GEM algorithm, $\mathcal{M}(\Psi^{(k)})$ is specified by the choice of $\Psi^{(k+1)}$ on the M-step so that

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) \geq Q(\Psi^{(k)}; \Psi^{(k)}).$$

Wu (1983) makes use of existing results in the optimization literature for point-to-set maps to establish conditions to ensure the convergence of a sequence of likelihood values $\{L(\Psi^{(k)})\}$ to a stationary value of $L(\Psi)$. The point-to-set map $\mathcal{M}(\Psi)$ is said to be closed at $\Psi = \Psi_o$ if $\Psi_m \rightarrow \Psi_o$, $\Psi_m \in \Omega$, and $\phi_m \rightarrow \phi_o$, $\phi_m \in \mathcal{M}(\Psi_m)$, implies that $\phi_o \in \mathcal{M}(\Psi_o)$. For a point-to-point map, continuity implies closedness.

Wu (1983) supposes that the following assumptions hold.

$$\Omega \text{ is a subset in } d\text{-dimensional Euclidean space } I\!\!R^d. \quad (3.18)$$

$$\Omega_{\Psi_o} = \{\Psi \in \Omega : L(\Psi) \geq L(\Psi_o)\} \text{ is compact for any } L(\Psi_o) > -\infty. \quad (3.19)$$

$$L(\Psi) \text{ is continuous in } \Omega \text{ and differentiable in the interior of } \Omega. \quad (3.20)$$

A consequence of the conditions (3.18) to (3.20) is that any sequence $\{L(\Psi^{(k)})\}$ is bounded above for any $\Psi^{(0)} \in \Omega$, where to avoid trivialities, it is assumed that the starting point $\Psi^{(0)}$ satisfies $L(\Psi^{(0)}) > -\infty$. As Wu (1983) acknowledges, the compactness assumption in (3.19) can be restrictive when no realistic compactification of the original parameter space is available. The regularity conditions (3.18) to (3.20) are assumed to hold for the remainder of this section and the next. It is also assumed there that each $\Psi^{(k)}$ is in the interior of Ω . That is, $\Psi^{(k+1)}$ is a solution of the equation

$$\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi = \mathbf{0}. \quad (3.21)$$

Wu (1983) notes that this assumption will be implied, for example, by

$$\Omega_{\Psi_o} \text{ is in the interior of } \Omega \text{ for any } \Psi_o \in \Omega. \quad (3.22)$$

An example where the compactness regularity condition (3.19) is not satisfied is a mixture of univariate normals with means μ_1 and μ_2 and unrestricted variances σ_1^2 and σ_2^2 . Each data point w_j gives rise to a singularity in $L(\Psi)$ on the edge of the parameter space Ω . More specifically, if $\mu_1(\mu_2)$ is set equal to w_j , then $L(\Psi)$ tends to infinity as $\sigma_1^2(\sigma_2^2)$ tends to zero. Thus in this example, if Ψ_o is any point in Ω with μ_1 or μ_2 set equal to w_j , then clearly Ω_{Ψ_o} is not compact. This space can be made compact by imposing the constraint $\sigma_i^2 \geq \epsilon$ ($i = 1, 2$). However, the condition (3.21) will not hold for EM sequences started sufficiently close to the boundary of the modified parameter space.

3.4.3 Main Convergence Theorem for a Generalized EM Sequence

We now state without proof the main convergence theorem given by Wu (1983) for a GEM algorithm. It also applies to the EM algorithm, since the latter is a special case of a GEM algorithm.

Theorem 3.1. Let $\{\Psi^{(k)}\}$ be an instance of a GEM algorithm generated by $\Psi^{(k+1)} \in \mathcal{M}(\Psi^{(k)})$. Suppose that (i) $\mathcal{M}(\Psi^{(k)})$ is closed over the complement of \mathcal{S} , the set of

stationary points in the interior of Ω and that (ii)

$$L(\Psi^{(k+1)}) > L(\Psi^{(k)}) \text{ for all } \Psi^{(k)} \notin \mathcal{S}.$$

Then all the limit points of $\{\Psi^{(k)}\}$ are stationary points and $L(\Psi^{(k)})$ converges monotonically to $L^* = L(\Psi^*)$ for some stationary point $\Psi^* \in \mathcal{S}$.

Condition (ii) holds for an EM sequence. For consider a $\Psi^{(k)} \notin \mathcal{S}$. Then from (3.15),

$$\begin{aligned} [\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi^{(k)}} &= \partial \log L(\Psi^{(k)}) / \partial \Psi \\ &\neq 0, \end{aligned}$$

since $\Psi^{(k)} \notin \mathcal{S}$. Hence $Q(\Psi; \Psi^{(k)})$ is not maximized at $\Psi = \Psi^{(k)}$, and so by the definition of the M-step of the EM algorithm and (3.11),

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) > Q(\Psi^{(k)}; \Psi^{(k)}).$$

This implies

$$L(\Psi^{(k+1)}) > L(\Psi^{(k)}),$$

thus establishing condition (ii) of Theorem 3.1.

For the EM algorithm, Wu (1983) notes that a sufficient condition for the closedness of the EM map is that

$$Q(\Psi; \phi) \text{ is continuous in both } \Psi \text{ and } \phi. \quad (3.23)$$

This condition is very weak and should hold in most practical situations. It has been shown to be satisfied for the case of a curved exponential family

$$g_c(\mathbf{x}; \Psi) = \exp\{\Psi^T \mathbf{t}(\mathbf{x}) - b(\Psi) + c(\mathbf{x})\},$$

where Ψ lies in a compact submanifold Ω_0 of the d -dimensional region Ω , as defined by (1.55).

This leads to the following theorem of Wu (1983) for an EM sequence.

3.4.4 A Convergence Theorem for an EM Sequence

Theorem 3.2. Suppose that $Q(\Psi; \phi)$ satisfies the continuity condition (3.23). Then all the limit points of any instance $\{\Psi^{(k)}\}$ of the EM algorithm are stationary points of $L(\Psi)$, and $L(\Psi^{(k)})$ converges monotonically to some value $L^* = L(\Psi^*)$ for some stationary point Ψ^* .

Theorem 3.2 follows immediately from Theorem 3.1, since its condition (i) is implied by the continuity assumption (3.23), while its condition (ii) is automatically satisfied for an EM sequence.

According to Wu (1983), Theorem 3.1 is the most general result for EM and GEM algorithms. Theorem 3.2 was obtained by Baum et al. (1970) and Haberman (1977) for two special models. Boyles (1983) gives a similar result for general models, but under stronger regularity conditions. One key condition in Baum et al. (1970) and Boyles (1983) is that $\mathcal{M}(\Psi^{(k)})$ is a continuous point-to-point map over Ω , which is stronger than a closed point-to-point map over the complement of \mathcal{S} , as assumed in Theorem 3.1.

From the viewpoint of the user, Theorem 3.2 provides the most useful result, since it only requires conditions that are easy to verify.

3.5 CONVERGENCE OF AN EM SEQUENCE OF ITERATES

3.5.1 Introduction

The convergence of the sequence of likelihood values $\{L(\Psi^{(k)})\}$ to some value L^* does not automatically imply the convergence of the corresponding sequence of iterates $\{\Psi^{(k)}\}$ to a point Ψ^* . But as Wu (1983) stresses, from a numerical viewpoint, the convergence of $\{\Psi^{(k)}\}$ is not as important as the convergence of $\{L(\Psi^{(k)})\}$ to a stationary value, in particular to a local maximum.

In this section, we present some results of Wu (1983) on the convergence of an EM sequence of iterates. As we shall see below, convergence of the EM iterates usually requires more stringent regularity conditions than for the convergence of the likelihood values.

3.5.2 Two Convergence Theorems of Wu (1983)

Define

$$\mathcal{S}(a) = \{\Psi \in \mathcal{S} : L(\Psi) = a\}$$

to be the subset of the set \mathcal{S} of stationary points in the interior of Ω at which $L(\Psi)$ equals a .

Theorem 3.3. Let $\{\Psi^{(k)}\}$ be an instance of a GEM algorithm satisfying conditions (i) and (ii) of Theorem 3.1. Suppose $\mathcal{S}(L^*)$ consists of the single point Ψ^* (that is, there cannot be two different stationary points with the same value L^*), where L^* is the limit of $L(\Psi^{(k)})$. Then $\Psi^{(k)}$ converges to Ψ^* .

This theorem follows immediately from Theorem 3.1 given that $\mathcal{S}(L^*)$ is a singleton. Wu (1983) notes that the above assumption $\mathcal{S}(L^*) = \{\Psi^*\}$ can be greatly relaxed if we assume

$$\|\Psi^{(k+1)} - \Psi^{(k)}\| \rightarrow 0, \text{ as } k \rightarrow \infty, \quad (3.24)$$

as a necessary condition for $\Psi^{(k)}$ to tend to Ψ^* .

Theorem 3.4. Let $\{\Psi^{(k)}\}$ be an instance of a GEM algorithm satisfying conditions (i) and (ii) of Theorem 3.1. If (3.24) holds, then all the limit points of $\{\Psi^{(k)}\}$ are in a connected and compact subset of $\mathcal{S}(L^*)$. In particular, if $\mathcal{S}(L^*)$ is discrete, then $\Psi^{(k)}$ converges to some Ψ^* in $\mathcal{S}(L^*)$.

Proof. From condition (3.19), $\{\Psi^{(k)}\}$ is a bounded sequence. According to Theorem 28.1 of Ostrowski (1966), the set of limit points of a bounded sequence $\{\Psi^{(k)}\}$ with (3.24) satisfied is connected and compact. From Theorem 3.1, all limit points of $\{\Psi^{(k)}\}$ are in $\mathcal{S}(L^*)$. \square

This theorem was proved by Boyles (1983) under different regularity conditions, while Theorem 3.2 was obtained by Hartley and Hocking (1971) for a special model.

Convergence of $\Psi^{(k)}$ to a stationary point can be proved without recourse to Theorem 3.1. For any value L_o , let

$$\mathcal{L}(L_o) = \{\Psi \in \Omega : L(\Psi) = L_o\}.$$

Theorem 3.5. Let $\{\Psi^{(k)}\}$ be an instance of a GEM algorithm with the additional property that

$$\partial[Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi^{(k+1)}} = \mathbf{0}. \quad (3.25)$$

Suppose $\partial Q(\Psi; \phi)/\partial\Psi$ is continuous in Ψ and ϕ . Then $\Psi^{(k)}$ converges to a stationary point Ψ^* with $L(\Psi^*) = L^*$, the limit of $L(\Psi^{(k)})$, if either

$$\mathcal{L}(L^*) = \{\Psi^*\} \quad (3.26)$$

or

$$\|\Psi^{(k+1)} - \Psi^{(k)}\| \rightarrow 0, \text{ as } k \rightarrow \infty, \text{ and } \mathcal{L}(L^*) \text{ is discrete.} \quad (3.27)$$

Proof. As noted earlier, the assumed regularity conditions (3.18) to (3.20) imply that $\{L(\Psi^{(k)})\}$ is bounded above, and so it converges to some point L^* . In the case of (3.26), where $\mathcal{L}(L^*)$ consists of the single point Ψ^* , $\Psi^{(k)}$ obviously converges to Ψ^* .

For $\mathcal{L}(L^*)$ discrete but not a singleton, condition (3.24) is sufficient to ensure that $\Psi^{(k)}$ converges to a point Ψ^* in $\mathcal{L}(L^*)$, as seen in the proof of Theorem 3.4.

From (3.15),

$$\partial \log L(\Psi^*) / \partial \Psi = [\partial Q(\Psi; \Psi^*) / \partial \Psi]_{\Psi=\Psi^*} \quad (3.28)$$

$$= \lim_{k \rightarrow \infty} [\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi^{(k+1)}} \quad (3.29)$$

$$= \mathbf{0}, \quad (3.30)$$

establishing that the limit point Ψ^* is a stationary point of $L(\Psi)$. In the above, (3.29) follows from (3.28) by the continuity of $\partial Q(\Psi; \phi)/\partial\Psi$ in Ψ and ϕ , while (3.30) follows from (3.29) by (3.25). \square

It should be noted that condition (3.25) is satisfied by any EM sequence under the regularity conditions assumed here. Since $\mathcal{S}(L)$ is a subset of $\mathcal{L}(L)$, conditions (3.26) and (3.27) of Theorem 3.5 are stronger than the corresponding ones in Theorems 3.3 and 3.4, respectively. The advantage of Theorem 3.5 is that it does not require the conditions (i) and (ii) of Theorem 3.1.

3.5.3 Convergence of an EM Sequence to a Unique Maximum Likelihood Estimate

An important special case of Theorem 3.5 in the previous subsection is the following corollary for a unimodal likelihood function $L(\Psi)$ with only one stationary point in the interior of Ω .

Corollary. Suppose that $L(\Psi)$ is unimodal in Ω with Ψ^* being the only stationary point and that $\partial Q(\Psi; \phi)\partial\Psi$ is continuous in Ψ and ϕ . Then any EM sequence $\{\Psi^{(k)}\}$ converges to the unique maximizer Ψ^* of $L(\Psi)$; that is, it converges to the unique MLE of Ψ .

3.5.4 Constrained Parameter Spaces

Wu's compactness condition is not satisfied, for instance, in the classic example of a mixture of univariate normals with means μ_1 and μ_2 and unrestricted variances σ_1^2 and σ_2^2 . Hathaway (1983, 1985) constrained the parameter space by placing inequality constraints on the variances and used the EM algorithm to obtain a consistent MLE. The convergence results established by Wu (1983) hold for the situation where the sequence of parameter estimates provided by the EM iteration lies entirely in the interior of the parameter space. In some estimation problems with constrained parameter spaces, the parameter value maximizing the log likelihood is on the boundary of the parameter space. Here some elements of the EM sequence may lie on the boundary, thus not fulfilling Wu's conditions for convergence.

Nettleton (1999) examines the behavior of the EM iterates in a number of problems with constrained parameter spaces; in some cases the EM sequence is guaranteed to converge to the MLE and in others not. He extends Wu's convergence results to the case of constrained parameter spaces and establishes some stricter conditions to guarantee convergence of the EM likelihood sequence to some local maximum and the EM parameter iterates to converge to the MLE. These conditions are on the EM sequence of iterates. One of his examples is an equal-proportion of mixture of a univariate standard normal and a univariate normal with mean $\theta \geq 0$ and unit variance. The parameter space here is constrained. He gives examples of data, where convergence is guaranteed and where it is not.

There are examples of effective use of the EM algorithm which are not covered by either Wu's or Nettleton's settings. One such example is Positron Emission Tomography (PET), where a spatial Poisson process model is used for image reconstruction using EM as the estimation technique (Vardi et al., 1985). This problem is discussed in Section 2.5. This is an example of the classical inversion problem or the system identification problem or the deconvolution problem, where a function is to be reconstructed from a collection of its line integrals or Radon transforms.

Here too, the EM iterates may fall on the boundary of the parameter space, but where Nettleton's conditions for guaranteed convergence are not satisfied. Special proofs of convergence seem to be needed for such cases as the ones provided by Lange and Carson (1984) and Iusem (1992).

Kim and Taylor (1995) describe an EM method when there are linear restrictions on the parameters.

3.6 EXAMPLES OF NONTYPICAL BEHAVIOR OF AN EM (GEM) SEQUENCE

3.6.1 Example 3.1: Convergence to a Saddle Point

In Example 2.1 in Section 2.2.1, we considered the fitting of a bivariate normal density to a random sample with some observations missing (completely at random) on each of the variates W_1 and W_2 . We now use this example in the particular case where the mean vector μ is specified to be zero to illustrate the possible convergence of the EM algorithm to a saddle point and not a local maximum. The data set taken from Murray (1977) consists of $n = 12$ pairs, four of which are complete, four incomplete without observations on Variate 1, and four incomplete without observations on Variate 2, making $m_1 + m_2 = 8$ incomplete pairs, as given below, where ? indicates a missing data point.

$$\begin{array}{cccccccccccc} \text{Variate 1:} & 1 & 1 & -1 & -1 & ? & ? & ? & ? & 2 & 2 & -2 & -2 \\ \text{Variate 2:} & 1 & -1 & 1 & -1 & 2 & 2 & -2 & -2 & ? & ? & ? & ? \end{array}$$

As μ is taken to be zero, the vector of unknown parameters is given by

$$\Psi = (\sigma_{11}, \sigma_{22}, \rho)^T.$$

In Figure 3.1, we have plotted $\log L(\Psi)$ against ρ and σ^2 under the imposition of the constraint

$$\sigma_{11} = \sigma_{22} = \sigma^2.$$

Murray (1977) reported that for the observed data y as listed above, the likelihood function $L(\Psi)$ has a saddle point at

$$\Psi_1 = (5/2, 5/2, 0)^T, \quad (3.31)$$

and two maxima at

$$\Psi_2 = (8/3, 8/3, \frac{1}{2})^T$$

and

$$\Psi_3 = (8/3, 8/3, -\frac{1}{2})^T.$$

He found that the EM sequence $\{\Psi^{(k)}\}$ converged to Ψ_1 if the EM algorithm were started from any point with $\rho = 0$. Otherwise, it converged to either Ψ_2 or Ψ_3 .

We now examine this further. From (2.5) with μ set equal to zero, the complete-data log likelihood is given by

$$\begin{aligned} \log L_c(\Psi) &= -n \log(2\pi) - \frac{1}{2}n \log \xi \\ &\quad - \frac{1}{2}\xi^{-1}(\sigma_{22}T_{11} + \sigma_{11}T_{22} - 2\sigma_{12}T_{12}) \end{aligned} \quad (3.32)$$

where

$$\xi = \sigma_{11}\sigma_{22}(1 - \rho^2)$$

and where the sufficient statistics T_{hi} ($h, i = 1, 2$) are defined by (2.7). To evaluate $Q(\Psi; \Psi^{(k)})$, the current conditional expectation of $\log L_c(\Psi)$ given \mathbf{y} , we need to calculate

$$E_{\Psi^{(k)}}(W_{ij} | \mathbf{y}) \text{ and } E_{\Psi^{(k)}}(W_{ij}^2 | \mathbf{y}).$$

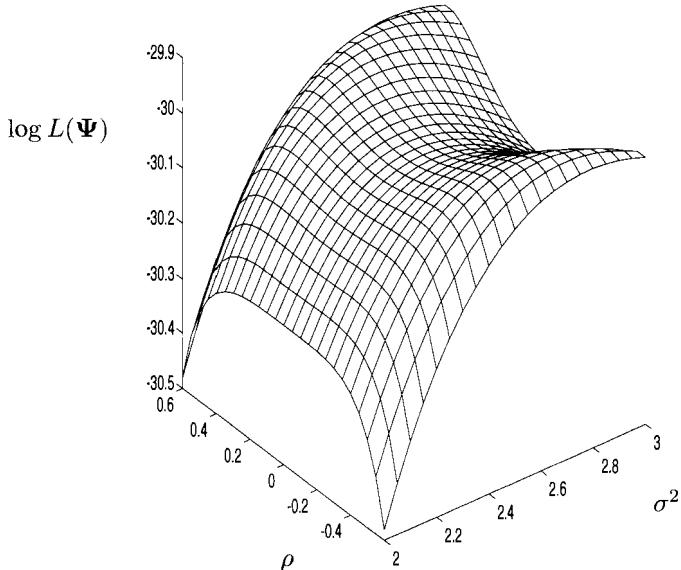


Figure 3.1 Plot of $\log L(\Psi)$ in the case $\sigma_{11} = \sigma_{22} = \sigma^2$.

From the previous results (2.10) and (2.12) on the conditional moments of the components of a normal random vector, we have for a missing w_{2j} that

$$E_{\Psi^{(k)}}(W_{2j} | \mathbf{y}) = w_{2j}^{(k)},$$

where

$$w_{2j}^{(k)} = \rho^{(k)} (\sigma_{22}^{(k)} / \sigma_{11}^{(k)})^{1/2} w_{1j}, \quad (3.33)$$

and

$$E_{\Psi^{(k)}}(W_{2j}^2 | \mathbf{y}) = w_{2j}^{(k)2} + \sigma_{22}^{(k)}(1 - \rho^{(k)2}). \quad (3.34)$$

The corresponding conditional moments for missing w_{1j} are obtained by interchanging the subscripts 1 and 2 in the right-hand side of (3.33) and of (3.34).

Let $\Psi_o^{(k)}$ be any value of $\Psi^{(k)}$ with $\rho^{(k)} = 0$. Then on using (3.33) and (3.34) and noting that

$$\sum_{j=1}^4 w_{1j} w_{2j} = 0,$$

it follows that

$$\begin{aligned} Q(\Psi; \Psi_o^{(k)}) &= -\frac{1}{2} \log(2\pi) - 6 \log(1 - \rho^2) \\ &\quad - 6 \log \sigma_{11} - 6 \log \sigma_{22} \\ &\quad - \frac{1}{2} \left(\sum_{j=1}^4 w_{1j}^2 + \sum_{j=9}^{12} w_{1j}^2 + 4\sigma_{11}^{(k)} \right) / \sigma_{11} \\ &\quad - \frac{1}{2} \left(\sum_{j=1}^8 w_{2j}^2 + 4\sigma_{22}^{(k)} \right) / \sigma_{22} \\ &= -\frac{1}{2} \log(2\pi) - 6 \log(1 - \rho^2) \\ &\quad - 6 \log \sigma_{11} - 6 \log \sigma_{22} \\ &\quad - (10 + 2\sigma_{11}^{(k)}) / \{\sigma_{11}(1 - \rho^2)\} \\ &\quad - (10 + 2\sigma_{22}^{(k)}) / \{\sigma_{22}(1 - \rho^2)\}. \end{aligned} \quad (3.35)$$

On equating the derivative of (3.35) with respect to Ψ to zero, it can be seen that the value $\Psi^{(k+1)}$ of Ψ that maximizes $Q(\Psi; \Psi_o^{(k)})$ is given by

$$\Psi^{(k+1)} = (\sigma_{11}^{(k+1)}, \sigma_{22}^{(k+1)}, \rho^{(k+1)})^T,$$

where

$$\rho^{(k+1)} = 0$$

and

$$\sigma_{ii}^{(k+1)} = (5 + \sigma_{ii}^{(k)}) / 3 \quad (i = 1, 2). \quad (3.36)$$

As $k \rightarrow \infty$, $\rho^{(k+1)}$ remains at zero, while on equating the right-hand side of (3.36) to $\sigma_{ii}^{(k+1)}$, we have that

$$\sigma_{ii}^{(k+1)} \rightarrow 5/2 \quad (i = 1, 2),$$

that is, $\Psi^{(k+1)}$ tends to Ψ_1 . This establishes the fact that the sequence $\{\Psi^{(k+1)}\}$ converges to Ψ_1 if started from any point with $\rho = 0$.

As noted earlier, the EM algorithm can converge to a saddle point Ψ_s of the (incomplete-data) log likelihood $L(\Psi)$ if $Q(\Psi; \Psi_s)$ is maximized at $\Psi = \Psi_s$. This is what happens in this example if the EM algorithm is started from any point with $\rho = 0$. It then converges to the saddle point $\Psi_s = \Psi_1$, as given by (3.31), with $Q(\Psi; \Psi_1)$ being maximized at $\Psi = \Psi_1$. In fact, $L(\Psi)$ is globally maximized at Ψ_1 when $\rho = 0$. However, if the EM algorithm is slightly perturbed from Ψ_1 , then it will diverge from Ψ_1 .

3.6.2 Example 3.2: Convergence to a Local Minimum

The fixed points of the EM algorithm include all local maxima of the likelihood, and sometimes saddle points and local minima. We have given an example where a fixed point is a saddle point. We now give an example from Arslan, Constable, and Kent (1993), where the EM algorithm can converge to a local minimum. When the likelihood has multiple local maxima, the parameter space can be partitioned into domains of convergence, one for each maximum. The prime reason for Arslan et al. (1993) presenting their examples was to demonstrate that these domains need not be connected sets.

For simplicity, Arslan et al. (1993) considered a special univariate case of the t -distribution (2.38) where only the location parameter μ is unknown and Σ (a scalar) is set equal to 1. In this special case, by taking the degrees of freedom ν very small, the likelihood function $L(\mu)$ can be made to have several local maxima.

From (2.47) and (2.48),

$$\mu^{(k+1)} = \sum_{j=1}^n u_j^{(k)} w_j \Bigg/ \sum_{j=1}^n u_j^{(k)},$$

where

$$u_j^{(k)} = \frac{\nu + 1}{\nu + (w_j - \mu^{(k)})^2}.$$

The mapping M induced here by the EM algorithm is therefore defined by

$$\mu^{(k+1)} = M(\mu^{(k)}),$$

where

$$M(\mu) = \sum_{j=1}^n u_j(\mu) w_j \Bigg/ \sum_{j=1}^n u_j(\mu)$$

and

$$u_j(\mu) = \frac{\nu + 1}{\nu + (w_j - \mu)^2} \quad (j = 1, \dots, n).$$

Arslan et al. (1993) took $\nu = 0.05$ and $n = 4$ and the observed data vector as $\mathbf{y} = (-20, 1, 2, 3)^T$. Thus ν is very small and one of the data points, $w_1 = -20$, might be regarded as an outlier.

The log likelihood function is plotted in Figure 3.2, where it can be seen that $\log L(\mu)$ has four local maxima at

$$\mu_1 = 19.993, \quad \mu_2 = 1.086, \quad \mu_3 = 1.997, \quad \text{and} \quad \mu_4 = 2.906,$$

and three local minima at

$$\mu_1^* = -14.516, \quad \mu_2^* = 1.373, \quad \text{and} \quad \mu_3^* = 2.647.$$

(Some of these values above as obtained in our analysis are slightly different from those reported in Arslan et al., 1993.)

The associated EM map $M(\mu)$ is plotted in Figure 3.3. It can be seen that it has seven fixed points as given above.

It can be seen from Figure 3.4, which is a blow-up of a portion of Figure 3.3, that there are two values of μ other than the fixed point μ_2^* for which

$$M(\mu) = \mu_2^*.$$

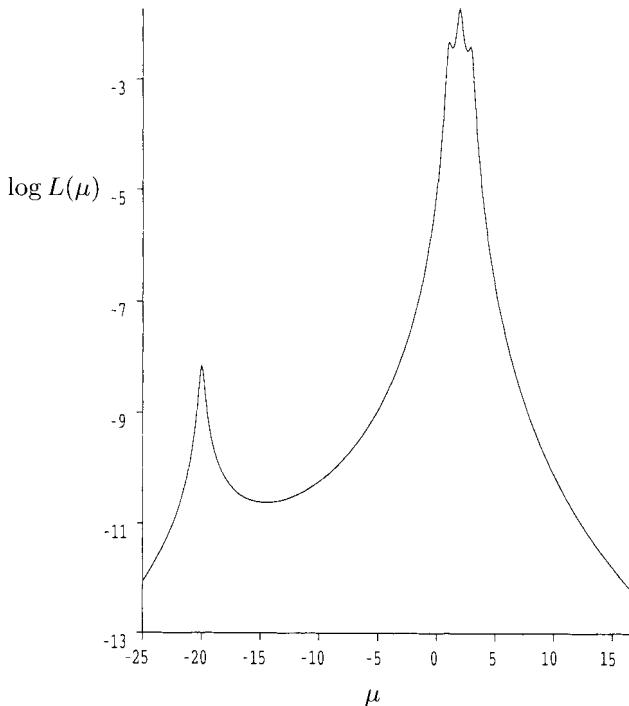


Figure 3.2 Plot of log likelihood function $\log L(\mu)$ versus μ .

We shall label these two points α_1 and α_2 , where $\alpha_1 = -1.874$ and $\alpha_2 = -0.330$. It follows that if the EM algorithm is started with $\mu^{(0)} = \alpha_1$ or α_2 , then

$$\begin{aligned}\mu^{(1)} &= M(\mu^{(0)}) \\ &= \mu_2^*,\end{aligned}\tag{3.37}$$

and so

$$\mu^{(k)} = \mu_2^* \quad (k = 1, 2, \dots).$$

That is, $\mu^{(k)}$ converges to the fixed point μ_2^* of the EM algorithm, corresponding to a local minimum of the likelihood function $L(\mu)$. Of course this curious behavior can happen only from these isolated starting points.

It is of interest to see here what happens if the EM algorithm is started from a point a long way from the data. As μ tends to $\pm\infty$,

$$u_j \approx \frac{\nu + 1}{\nu + \mu^2}$$

for all j , and so

$$\begin{aligned}M(\mu) &\approx \sum_{j=1}^n w_j/n \\ &= \bar{w}.\end{aligned}\tag{3.38}$$

Thus when the starting value $\mu^{(0)}$ is taken very far away from the data, $\mu^{(1)}$ is somewhere near the sample mean.

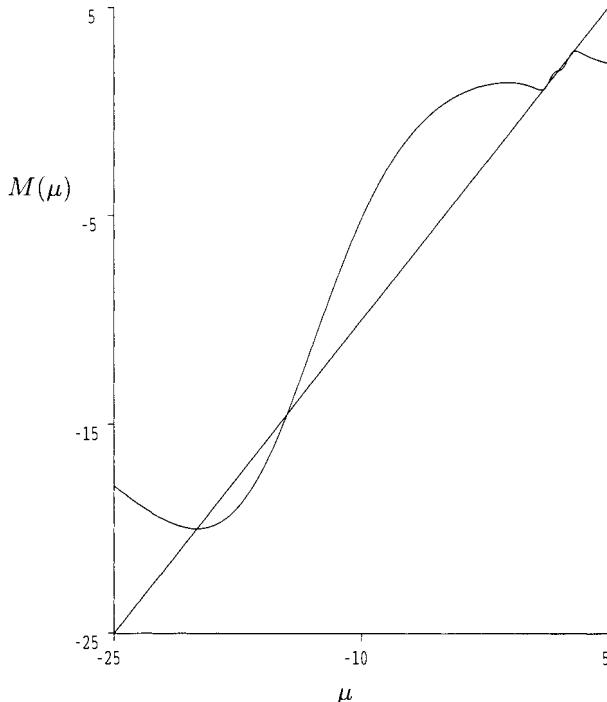


Figure 3.3 Plot of EM mapping $M(\mu)$ versus μ .

It should be stressed that since ν is very small, the example given here is artificial. However, as argued by Arslan et al. (1993), it provides a useful warning because estimation for the one-parameter t -distribution is so straightforward. If erratic convergence behavior can occur in this example, who knows what pitfalls there may be when the EM algorithm is used in more complicated settings where multiple maxima are present.

3.6.3 Example 3.3: Nonconvergence of a Generalized EM Sequence

We now present an example of a GEM sequence for which $L(\Psi^{(k)})$ converges, but for which $\Psi^{(k)}$ does not converge to a single point. This example was originally presented by Boyles (1983) as a counterexample to a claim incorrectly made in Theorem 2 of DLR that, under their conditions (1) and (2), a GEM sequence $\{\Psi^{(k)}\}$ converges to a point Ψ^* in the closure of Ω .

Suppose that $\phi(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the bivariate normal density with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and known covariance matrix $\boldsymbol{\Sigma}$ taken equal to the (2×2) identity matrix \mathbf{I}_2 . In this example of Boyles (1983), there are no missing data, so that

$$\mathbf{y} = \mathbf{x} = (w_1, w_2)^T,$$

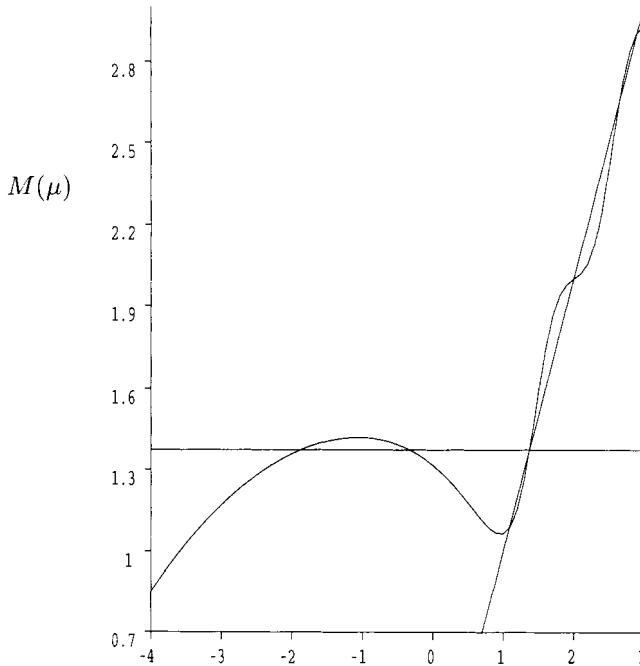


Figure 3.4 Blow-up of Figure 3.3.

where x consists of the single observation $\mathbf{w} = (w_1, w_2)^T$. The vector Ψ of unknown parameters is μ , and

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) &= \log L(\mu) \\ &= \log \phi(\mathbf{w}; \mu, \mathbf{I}_2) \\ &= (2\pi)^{-1} \exp\left\{-\frac{1}{2}(\mathbf{w} - \mu)^T(\mathbf{w} - \mu)\right\}. \end{aligned}$$

The GEM sequence is defined to be

$$\mu_1^{(k)} = w_1 + r^{(k)} \cos \theta^{(k)}$$

and

$$\mu_2^{(k)} = w_2 + r^{(k)} \sin \theta^{(k)},$$

where

$$r^{(k)} = 1 + (k + 1)^{-1}$$

and

$$\theta^{(k)} = \sum_{i=1}^k (i + 1)^{-1}$$

for $k = 1, 2, \dots$ and where r and θ are the polar coordinates centered at the single observation \mathbf{w} . The initial value for r and θ are $r^{(0)} = 2$ and $\theta^{(0)} = 0$.

We have that

$$\begin{aligned}\log L(\Psi^{(k+1)}) - \log L(\Psi^{(k)}) &= \frac{1}{2}(r^{(k)2} - r^{(k+1)2}) \\ &= \frac{1}{2}\{r^{(k)2} - (2 - \frac{1}{r^{(k)2}})^2\},\end{aligned}\quad (3.39)$$

since

$$r^{(k+1)} = 2 - \frac{1}{r^{(k)}}.$$

Now

$$0 < 2 - u^{-1} \leq u \quad (3.40)$$

for $u \geq 1$. As $r^{(k)} \geq 1$ for each k , we have from (3.39) and (3.40) that

$$L(\Psi^{(k+1)}) - L(\Psi^{(k)}) \geq 0.$$

Hence $\{\Psi^{(k)}\}$ is an instance of a GEM algorithm.

Since $r^{(k)} \rightarrow 1$, as $k \rightarrow \infty$, the sequence of likelihood values $\{L(\Psi^{(k)})\}$ converges to the value

$$(2\pi)^{-1}e^{-\frac{1}{2}}.$$

But the sequence of iterates $\{\Psi^{(k)}\}$ converges to the circle of unit radius and center w . It does converge not to a single point, although it does satisfy the conditions of Theorem 2 of DLR for the convergence of a GEM sequence to a single point. These conditions are that $\{L(\Psi^{(k)})\}$ is a bounded sequence, which is obviously satisfied here, and that

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) - Q(\Psi^{(k)}; \Psi^{(k)}) \geq \lambda \|\Psi^{(k+1)} - \Psi^{(k)}\|^2 \quad (3.41)$$

for some scalar $\lambda > 0$ and all k . Now the left-hand side of (3.41) equals

$$\frac{1}{2}(r^{(k)2} - r^{(k+1)2})$$

while, concerning the right-hand side of (3.41),

$$\|\Psi^{(k+1)} - \Psi^{(k)}\|^2 = r^{(k)2} + r^{(k+1)2} - 2r^{(k)}r^{(k+1)}\cos((k+2)^{-1}).$$

Taking $\lambda = \frac{1}{2}$, the inequality (3.41) holds if

$$r^{(k+1)}\{r^{(k)}\cos((k+2)^{-1}) - r^{(k+1)}\} \geq 0,$$

which is equivalent to

$$\cos((k+2)^{-1}) \geq 1 - (k+2)^{-2},$$

and the latter is valid for all $k \geq 0$.

Boyles (1983) notes that although these conditions do not imply the convergence of $\{\Psi^{(k)}\}$, they do imply that

$$\|\Psi^{(k+1)} - \Psi^{(k)}\| \rightarrow 0, \text{ as } k \rightarrow \infty.$$

But as we have seen in this section, this latter condition is not sufficient to ensure convergence to a single point.

3.6.4 Example 3.4: Some E-Step Pathologies

Example 2.4 of Section 2.4 may have given an impression that the E-step consists in replacing the missing data by their conditional expectations given the observed data at current parameter values. Although in many examples this may be the case, it is not quite so in general; rather, as should be clear from the general theory described above, the E-step consists in replacing the log likelihood by its conditional expectation given the observed data at current parameter values. In the example with missing data from the bivariate normal, discussed in Section 2.2.1, we noticed that not only the missing data z is to be replaced in the E-step by its conditional expectation, but also z^2 the replacement value of which is its conditional expectation, which is not the square of the conditional expectation of z . Flury and Zoppé (2000) give the following interesting example to demonstrate this point that the E-step does not always consist in plugging in “estimates” for missing data. This is also an example where the likelihood vanishes at a point in the parameter space during the course of the E-step leading to a difficulty in computing this step.

Let $c > 0$ be a fixed time point. Let the observed data \mathbf{y} consist of $n + m$ independent observations w_1, \dots, w_n and $\delta_{n+1}, \dots, \delta_{n+m}$, where the w_j 's are exact lifetimes of a random sample of n bulbs and δ_j 's are indicator observations on an independent random sample of m bulbs, with $\delta_j = 1$ if the j th bulb is still burning at time c ($w_j > c$) and 0 otherwise. Let m_c be the number of δ_j 's with value 1, and let $w_{\max} = \max\{w_1, \dots, w_n\}$.

Let us first work out the MLE of θ directly on the basis of the observed data \mathbf{y} . The likelihood function for θ is

$$L(\theta) = \theta^{-n} \left(\frac{c}{\max(c, \theta)} \right)^{m-m_c} \left(1 - \frac{c}{\max(c, \theta)} \right)^{m_c} I_{[w_{\max}, \infty)}(\theta). \quad (3.42)$$

For $m_c = 0$, $L(\theta)$ is decreasing in θ for $\theta \geq w_{\max}$ and hence the MLE is $\hat{\theta} = w_{\max}$. For $m_c \geq 1$, we have $\theta \geq c$. Now $L(\theta) = \text{constant} \times L_1(\theta)$, where $L_1(\theta) = \theta^{-(n+m)} (\theta - c)^{m_c}$, which has a unique maximum at $\hat{\theta} = \frac{n+m}{n+m-m_c} c$ and is monotonically decreasing for $\theta > \hat{\theta}$. Hence the MLE of θ is

$$\hat{\theta} = \begin{cases} \hat{\theta} & \text{if } \hat{\theta} > w_{\max} \text{ and } m_c \geq 1, \\ w_{\max} & \text{otherwise.} \end{cases}$$

Now let us try the EM algorithm for the case $m_c \geq 1$. The complete-data for this case may be formulated as $w_1, \dots, w_n, w_{n+1}, w_{n+2}, \dots, w_{n+m}$ where the last m are the actual lifetimes of the second sample of m bulbs. The complete-data MLE of θ is w_{\max} . Since $m_c \geq 1$, we have $\theta \geq c$. Now if we take the approach of replacing the missing (censored) observations w_j by their (current) conditional expectations given the observed data \mathbf{y} , then we simply need the result that

$$E_{\theta^{(k)}}(W_j | \delta_j = 1) = \frac{1}{2}(c + \theta^{(k)}) \quad (j = n+1, \dots, n+m).$$

The M-step is

$$\theta^{(k+1)} = \max\{w_{\max}, \frac{1}{2}(c + \theta^{(k)})\}.$$

Combining the E- and M-steps, we can write the EM algorithm as a sequence of iterations of the equation

$$\theta^{(k+1)} = M(\theta^{(k)}) \equiv \max\{w_{\max}, \frac{1}{2}(c + \theta^{(k)})\}.$$

It can be seen that if we start with any $\theta^{(0)}$, this procedure will converge to $\hat{\theta} = \max\{w_{\max}, c\}$, by noting that this $\hat{\theta}$ satisfies the self-consistency equation $\theta = M(\theta)$. Thus plugging-in

expected values of missing values is not necessarily the E-step of the EM algorithm and does not necessarily lead to MLEs.

In the E-step, we are supposed to find the conditional expectation of complete-data log likelihood given y , at the current value of the parameter, which is not always the same operation as imputation of missing values. Now given the data with $m_c \geq 1$, we have $\theta \geq c$ and hence the conditional distributions of w_j are uniform in $[c, \theta^{(k)}]$. Thus for $\theta < \theta^{(k)}$ the conditional density of missing w_j takes value 0 with positive probability and hence the complete-data log likelihood and its conditional expected value we are seeking do not exist; this can be seen from (3.42). Thus Flury and Zoppé (2000) argue that the E-step cannot be carried out and that the EM algorithm is not applicable.

However, Hunter (2003) shows that a legitimate EM algorithm can indeed be constructed, although it does not result in the MLE. Since θ should be at least as large as $\ell = \max\{w_{\max}, c\}$, let us start with $\theta^{(0)} > \ell$. Any w_{n+j} value for $j = 1, \dots, m$ has a uniform distribution on $[c, \theta^{(0)}]$ and hence if $\theta < \theta^{(0)}$, then $w_{\max} > \theta$ has a positive probability; thus the conditional expectation of log likelihood is

$$Q(\theta; \theta^{(0)}) = \begin{cases} -(n+m) \log \theta & \text{if } \theta \geq \theta^{(0)}, \\ -\infty & \text{if } 0 < \theta < \theta^{(0)}. \end{cases} \quad (3.43)$$

Noticing that Q is decreasing on (θ_0, ∞) and $Q(\theta; \theta^{(0)}) < Q(\theta^{(0)}; \theta^{(0)})$ on $(0, \theta^{(0)})$, the maximization of (3.43) gives $\theta^{(1)} = \theta^{(0)}$ and hence the EM algorithm is stuck at $\theta^{(0)}$, though it is not the MLE. Hunter (2003) points out that this situation is due to the non-differentiability of the Q function at $\theta = \theta^{(0)}$.

Hunter (2003) goes on to construct an example in which the EM algorithm converges to the correct MLE despite the likelihood vanishing at a point in the parameter space. Here n lifetimes are from uniform $(0, \theta)$ and m are from exponential with mean θ , censored at time c . Using the notation $\bar{w} = \frac{1}{m} \sum_{i=n+1}^{n+m} w_j$, we have:

$$\log L_c(\theta) = \begin{cases} -(n+m) \log \theta - m\bar{w}/\theta & \text{if } \theta \geq w_{\max}, \\ -\infty & \text{if } 0 < \theta < w_{\max}, \end{cases}$$

$$\log L(\theta) = \begin{cases} -n \log \theta + (m - m_c) \log(1 - e^{-c/\theta}) - m_c c / \theta & \text{if } \theta \geq w_{\max}, \\ -\infty & \text{if } 0 < \theta < w_{\max}, \end{cases}$$

$$Q(\theta; \theta^{(k)}) = \begin{cases} -(n+m) \log \theta - e_k / \theta & \text{if } \theta \geq w_{\max}, \\ -\infty & \text{if } 0 < \theta < w_{\max}, \end{cases}$$

where

$$e_k = m_c c + m\theta^{(k)} + \frac{c(m - m_c)}{1 - e^{c/\theta^{(k)}}}.$$

If we start at any $\theta^{(0)} > w_{\max}$, then the EM algorithm results in

$$\theta^{(k+1)} = \begin{cases} e_k / (m+n) & \text{if } w_{\max} < e_k / (m+n), \\ w_{\max} & \text{otherwise.} \end{cases}$$

Much of Hunter's insight resulting in these examples is gleaned from the geometry of the EM algorithm and its generalized version, the MM algorithm; see Section 7.7 for some details of the MM algorithm and its geometry.

3.7 SCORE STATISTIC

We let

$$\mathbf{S}(\mathbf{y}; \boldsymbol{\Psi}) = \partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi}$$

be the gradient vector of the log likelihood function $L(\boldsymbol{\Psi})$; that is, the score statistic based on the observed (incomplete) data \mathbf{y} . The gradient vector of the complete-data log likelihood function is given by

$$\mathbf{S}_c(\mathbf{x}; \boldsymbol{\Psi}) = \partial \log L_c(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi}.$$

The incomplete-data score statistic $\mathbf{S}(\mathbf{y}; \boldsymbol{\Psi})$ can be expressed as the conditional expectation of the complete-data score statistic; that is,

$$\mathbf{S}(\mathbf{y}; \boldsymbol{\Psi}) = E_{\boldsymbol{\Psi}}\{\mathbf{S}_c(\mathbf{X}; \boldsymbol{\Psi}) \mid \mathbf{y}\}. \quad (3.44)$$

To see this, we note that

$$\begin{aligned} \mathbf{S}(\mathbf{y}; \boldsymbol{\Psi}) &= \partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} \\ &= \partial \log g(\mathbf{y}; \boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} \\ &= g'(\mathbf{y}; \boldsymbol{\Psi}) / g(\mathbf{y}; \boldsymbol{\Psi}) \\ &= \left\{ \int_{\mathcal{X}(\mathbf{y})} g'_c(\mathbf{x}; \boldsymbol{\Psi}) d\mathbf{x} \right\} / g(\mathbf{y}; \boldsymbol{\Psi}), \end{aligned} \quad (3.45)$$

where the prime denotes differentiation with respect to $\boldsymbol{\Psi}$. On multiplying and dividing the integrand by $g_c(\mathbf{x}; \boldsymbol{\Psi})$ in (3.45), we have that

$$\begin{aligned} \mathbf{S}(\mathbf{y}; \boldsymbol{\Psi}) &= \int_{\mathcal{X}(\mathbf{y})} \{\partial \log g_c(\mathbf{x}; \boldsymbol{\Psi}) / \partial \boldsymbol{\Psi}\} \{g_c(\mathbf{x}; \boldsymbol{\Psi}) / g(\mathbf{y}; \boldsymbol{\Psi})\} d\mathbf{x} \\ &= \int_{\mathcal{X}(\mathbf{y})} \{\partial \log L_c(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi}\} k(\mathbf{x} \mid \mathbf{y}; \boldsymbol{\Psi}) d\mathbf{x} \\ &= E_{\boldsymbol{\Psi}}\{\partial \log L_c(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} \mid \mathbf{y}\} \\ &= E_{\boldsymbol{\Psi}}\{\mathbf{S}_c(\mathbf{X}; \boldsymbol{\Psi}) \mid \mathbf{y}\}. \end{aligned} \quad (3.46)$$

Another way of writing this result is that

$$\mathbf{S}(\mathbf{y}; \boldsymbol{\Psi}) = [\partial Q(\boldsymbol{\Psi}_o; \boldsymbol{\Psi}) / \partial \boldsymbol{\Psi}_o]_{\boldsymbol{\Psi}_o=\boldsymbol{\Psi}}, \quad (3.47)$$

which is the self-consistency property of the EM algorithm, as noted in Section 3.2. The result (3.47) follows from (3.46) on interchanging the operations of expectation and differentiation. It is assumed in the above that regularity conditions hold for this interchange.

Note that the result (3.47) was derived by an alternative argument in Section 3.4; see equation (3.15).

3.8 MISSING INFORMATION

3.8.1 Missing Information Principle

We let

$$\mathbf{I}(\boldsymbol{\Psi}; \mathbf{y}) = -\partial^2 \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T$$

be the matrix of the negative of the second-order partial derivatives of the (incomplete-data) log likelihood function with respect to the elements of Ψ . Under regularity conditions, the expected (Fisher) information matrix $\mathcal{I}(\Psi)$ is given by

$$\begin{aligned}\mathcal{I}(\Psi) &= E_{\Psi}\{S(\mathbf{Y}; \Psi)S^T(\mathbf{Y}; \Psi)\} \\ &= E_{\Psi}\{\mathbf{I}(\Psi; \mathbf{Y})\}.\end{aligned}$$

With respect to the complete-data log likelihood, we let

$$\mathbf{I}_c(\Psi; \mathbf{x}) = -\partial^2 \log L_c(\Psi)/\partial\Psi\partial\Psi^T.$$

The expected (complete-data) information matrix is given then by

$$\mathcal{I}_c(\Psi) = -E_{\Psi}\{\mathbf{I}_c(\Psi; \mathbf{X})\}.$$

We have from (3.3) that

$$\log L(\Psi) = \log L_c(\Psi) - \log k(\mathbf{x} | \mathbf{y}; \Psi). \quad (3.48)$$

On differentiating the negative of both sides of (3.48) twice with respect to Ψ , we have that

$$\mathbf{I}(\Psi; \mathbf{y}) = \mathbf{I}_c(\Psi; \mathbf{x}) + \partial^2 \log k(\mathbf{x} | \mathbf{y}; \Psi)/\partial\Psi\partial\Psi^T. \quad (3.49)$$

Taking the expectation of both sides of (3.49) over the conditional distribution of \mathbf{z} given \mathbf{y} yields

$$\mathbf{I}(\Psi; \mathbf{y}) = \mathcal{I}_c(\Psi; \mathbf{y}) - \mathcal{I}_m(\Psi; \mathbf{y}), \quad (3.50)$$

where

$$\mathcal{I}_c(\Psi; \mathbf{y}) = E_{\Psi}\{\mathbf{I}_c(\Psi; \mathbf{X}) | \mathbf{y}\} \quad (3.51)$$

is the conditional expectation of the complete-data information matrix $\mathbf{I}_c(\Psi; \mathbf{X})$ given \mathbf{y} , and where

$$\mathcal{I}_m(\Psi; \mathbf{y}) = -E_{\Psi}\{\partial^2 \log k(\mathbf{X} | \mathbf{y}; \Psi)/\partial\Psi\partial\Psi^T | \mathbf{y}\} \quad (3.52)$$

is the expected information matrix for Ψ based on \mathbf{x} (or equivalently, the unobservable data \mathbf{z}) when conditioned on \mathbf{y} . This latter information denoted by $\mathcal{I}_m(\Psi; \mathbf{y})$ can be viewed as the “missing information” as a consequence of observing only \mathbf{y} and not also \mathbf{z} . Thus (3.50) has the following interpretation: observed information equals the (conditional expected) complete information minus the missing information. This has been called the Missing Information principle by Orchard and Woodbury (1972).

On averaging both sides of (3.50) over the distribution of \mathbf{Y} , we have

$$\mathcal{I}(\Psi) = \mathcal{I}_c(\Psi) - E_{\Psi}\{\mathcal{I}_m(\Psi; \mathbf{Y})\}$$

as an analogous expression to (3.50) for the expected information $\mathcal{I}(\Psi)$. Orchard and Woodbury (1972) give a slightly different form of this expression.

3.8.2 Example 3.5: Censored Exponentially Distributed Survival Times (Example 1.3 Continued)

To demonstrate the relationship (3.50) between the incomplete-data, complete-data, and missing information matrices, we return to Example 1.3 on censored exponentially distributed survival times.

From (3.50), we have

$$I(\mu; \mathbf{y}) = I_c(\mu; \mathbf{y}) - I_m(\mu; \mathbf{y}), \quad (3.53)$$

where $I(\mu; \mathbf{y})$ is the information about μ in the observed data \mathbf{y} , $I_c(\mu; \mathbf{y})$ is the conditional expectation of the complete-data information, and $I_m(\mu; \mathbf{y})$ is the missing information. For this example, we shall show that this relationship holds by calculating these three information quantities from their definitions.

By differentiation of the incomplete-data log likelihood function (1.46), we have that

$$I(\mu; \mathbf{y}) = -r\mu^{-2} + 2\mu^3 \sum_{j=1}^n c_j. \quad (3.54)$$

Concerning $I_c(\mu; \mathbf{y})$, we have on differentiation of minus the complete-data log likelihood twice with respect to μ , that

$$I_c(\mu; \mathbf{x}) = (-n\mu^{-2} + 2\mu^{-3} \sum_{j=1}^n w_j),$$

and so

$$\begin{aligned} I_c(\mu; \mathbf{y}) &= E_\mu\{I_c(\mu; \mathbf{x}) \mid \mathbf{y}\} \\ &= \{-n\mu^{-2} + 2\mu^{-3} \sum_{j=1}^r c_j \\ &\quad + 2\mu^{-3} \sum_{j=r+1}^n (c_j + \mu)\} \\ &= (n - 2r)\mu^{-2} + 2\mu^{-3} \sum_{j=1}^n c_j, \end{aligned} \quad (3.55)$$

using (1.50).

The missing information $I_m(\mu; \mathbf{y})$ can be calculated directly from its definition (3.52),

$$I_m(\mu; \mathbf{y}) = -E_\mu\{\partial^2 \log k(\mathbf{X} \mid \mathbf{y}; \mu) / \partial \mu^2 \mid \mathbf{y}\}.$$

Given \mathbf{y} , the conditional joint distribution of W_1, \dots, W_r is degenerate with mass one at the point $(c_1, \dots, c_r)^T$. Hence we can effectively work with the conditional distribution of \mathbf{Z} given \mathbf{y} in calculating $I_m(\mu; \mathbf{y})$ from (3.52). From (1.49), the conditional density of \mathbf{Z} given \mathbf{y} is

$$k(\mathbf{z} \mid \mathbf{y}; \mu) = \mu^{-(n-r)} \exp\left\{-\frac{1}{2} \sum_{j=r+1}^n (w_j - c_j)/\mu\right\} I_A(\mathbf{z}), \quad (3.56)$$

where

$$A = \{\mathbf{z} : w_j > c_j \ (j = r + 1, \dots, n)\}.$$

On differentiation of (3.56), we have that

$$\begin{aligned} I_m(\mu; \mathbf{y}) &= -(n - r)\mu^{-2} + 2\mu^{-3} E_\mu\left\{\sum_{j=r+1}^n (W_j - c_j) \mid \mathbf{y}\right\} \\ &= -(n - r)\mu^{-2} + 2\mu^{-3} \sum_{j=r+1}^n \mu \\ &= (n - r)\mu^{-2}, \end{aligned} \quad (3.57)$$

since from (1.50),

$$E_\mu\{(W_j - c_j) \mid \mathbf{y}\} = \mu$$

for $j = r + 1, \dots, n$. On subtracting (3.57) from (3.55), we confirm the expression (3.54) for $I(\mu; \mathbf{y})$ obtained directly from its definition.

In Section 4.2, it will be seen that $I_m(\mu; \mathbf{y})$ can be expressed as

$$I_m(\mu; \mathbf{y}) = \text{var}_\mu\{S_c(\mathbf{X}; \mu) \mid \mathbf{y}\}. \quad (3.58)$$

On using this definition, we have

$$I_m(\mu; \mathbf{y}) = \text{var}_\mu\{S_c(\mathbf{X}; \mu) \mid \mathbf{y}\} \quad (3.59)$$

$$\begin{aligned} &= \text{var}_\mu\{\mu^{-2} \sum_{j=1}^n (W_j - c_j) \mid \mathbf{y}\} \\ &= (n - r)\mu^{-2}, \end{aligned} \quad (3.60)$$

since from (1.49), it follows that

$$\begin{aligned} \text{var}_\mu(W_j \mid \mathbf{y}) &= 0 \quad (j = 1, \dots, r), \\ &= \mu^2 \quad (j = r + 1, \dots, n). \end{aligned}$$

It can be seen that the result (3.60) agrees with (3.57) as obtained directly from its definition.

In order to compute the expected information $I(\mu)$, we have to make some assumption about the underlying censoring mechanism. Suppose that an observation W_j is censored if its realized value w_j exceeds C . That is, an observation is censored if it fails after time C from when the experiment was commenced. Then its expected information $I(\mu)$ is given by

$$\begin{aligned} I(\mu) &= E_\mu\{I(\mu; \mathbf{y})\} \\ &= E_\mu\{-r\mu^{-2} + 2\mu^{-3} \sum_{j=1}^n W_j\} \\ &= \mu^{-2} E_\mu\{-r + 2\mu^{-1} \sum_{j=1}^n W_j\}. \end{aligned} \quad (3.61)$$

Now under the assumed censoring scheme,

$$\begin{aligned} E_\mu(r) &= n \Pr\{W_j < C\} \\ &= n(1 - e^{-C/\mu}), \end{aligned} \quad (3.62)$$

and

$$E_\mu(W_j) = \mu(1 - e^{-C/\mu}). \quad (3.63)$$

Thus

$$I(\mu) = n\mu^{-2}(1 - e^{-C/\mu}). \quad (3.64)$$

As the complete-data expected information $I_c(\mu)$ is $n\mu^{-2}$, it can be seen that the expected missing information about μ is

$$E_\mu\{I_m(\mu; \mathbf{Y})\} = n\mu^{-2}e^{-C/\mu}, \quad (3.65)$$

which tends to zero, as $C \rightarrow \infty$. That is, as C becomes large, most items will fail before this censoring time, and so the amount of missing information due to censoring becomes negligible.

For this example, the complete-data density belongs to the regular exponential family with natural parameter $\theta = \mu^{-1}$ and sufficient statistic

$$t(\mathbf{X}) = \sum_{j=1}^n W_j.$$

Hence if we work directly in terms of the information about θ , rather than μ , verification of (3.50) is trivial. From from (1.58) and (1.59),

$$\begin{aligned} I_c(\theta; \mathbf{y}) &= \text{var}_\theta\{t(\mathbf{X})\} \\ &= n\theta^2, \end{aligned}$$

while from (3.59),

$$\begin{aligned} I_m(\theta; \mathbf{y}) &= \text{var}_\theta\{t(\mathbf{X}) \mid \mathbf{y}\} \\ &= (n - r)\theta^2, \end{aligned}$$

thus establishing

$$\begin{aligned} I(\theta; \mathbf{y}) &= r\theta^2 \\ &= r\mu^{-2}. \end{aligned}$$

To illustrate this, Hunter (2003) has constructed an example involving the estimation of an exponential mean μ . The data consists of n uncensored observations on the distribution with mean \bar{y} and m censored observations for which we only know if they exceed a given number t ; let z be the number that exceed t . He constructs two situations one with more censored (missing) data and another with less, such that the MLE of μ is the same. The values of s defined above for the two cases are given in the following table, which shows how more missing data has slower speed of convergence.

Table 3.1 Missing Data and Speed of Convergence.

Missing Data	m	n	t	z	\bar{y}	$\hat{\mu}$	Speed s
More	18	2	2	7	1.401745	2	0.6080
Less	2	18	2	1	1.953553	2	0.9462

Source: Hunter (2003)

3.9 RATE OF CONVERGENCE OF THE EM ALGORITHM

3.9.1 Rate Matrix for Linear Convergence

We have seen in Section 1.5 that the EM algorithm implicitly defines a mapping $\Psi \rightarrow \mathcal{M}(\Psi)$, from the parameter space of Ψ , Ω , to itself such that each iteration $\Psi^{(k)} \rightarrow \Psi^{(k+1)}$ is defined by

$$\Psi^{(k+1)} = \mathcal{M}(\Psi^{(k)}) \quad (k = 0, 1, 2, \dots).$$

If $\Psi^{(k)}$ converges to some point Ψ^* and $M(\Psi)$ is continuous, then Ψ^* is a fixed point of the algorithm; that is, Ψ^* must satisfy

$$\Psi^* = M(\Psi^*). \quad (3.66)$$

By a Taylor series expansion of $\Psi^{(k+1)} = M(\Psi^{(k)})$ about the point $\Psi^{(k)} = \Psi^*$, we have in a neighborhood of Ψ^* that

$$\Psi^{(k+1)} - \Psi^* \approx J(\Psi^*)(\Psi^{(k)} - \Psi^*), \quad (3.67)$$

where $J(\Psi)$ is the $d \times d$ Jacobian matrix for $M(\Psi) = (M_1(\Psi), \dots, M_d(\Psi))^T$, having (i, j) th element $J_{ij}(\Psi)$ equal to

$$J_{ij}(\Psi) = \partial M_i(\Psi) / \partial \Psi_j, \quad (3.68)$$

where $\Psi_j = (\Psi)_j$. Note that $J(\Psi)$ is the transpose of the matrix DM that is used by DLR and Meng and Rubin (1991).

Thus, in a neighborhood of Ψ^* , the EM algorithm is essentially a linear iteration with rate matrix $J(\Psi^*)$, since $J(\Psi^*)$ is typically nonzero. For this reason, $J(\Psi^*)$ is often referred to as the matrix rate of convergence, or simply, the rate of convergence.

For vector Ψ , a measure of the actual observed convergence rate is the global rate of convergence, which is defined as

$$r = \lim_{k \rightarrow \infty} \|\Psi^{(k+1)} - \Psi^*\| / \|\Psi^{(k)} - \Psi^*\|,$$

where $\|\cdot\|$ is any norm on d -dimensional Euclidean space \mathbb{R}^d . It is well known that under certain regularity conditions,

$$r = \lambda_{\max} \equiv \text{the largest eigenvalue of } J(\Psi^*).$$

Notice that a large value of r implies slow convergence. To be consistent with the common notion that the higher the value of the measure, the faster the algorithm converges, Meng (1994) defines $s = 1 - r$ as the global speed of convergence. Thus s is the smallest eigenvalue of

$$S = I_d - J(\Psi^*), \quad (3.69)$$

which may be called the (matrix) speed of convergence; S is often referred to as the iteration matrix in the optimization literature.

3.9.2 Measuring the Linear Rate of Convergence

In practice, r is typically assessed as

$$r = \lim_{k \rightarrow \infty} \|\Psi^{(k+1)} - \Psi^{(k)}\| / \|\Psi^{(k)} - \Psi^{(k-1)}\|. \quad (3.70)$$

The rate of convergence can also be measured component by component. The i th componentwise rate of convergence is defined as

$$r_i = \lim_{k \rightarrow \infty} |\Psi_i^{(k+1)} - \Psi_i^*| / |\Psi_i^{(k)} - \Psi_i^*|,$$

provided that it exists. We define $r_i \equiv 0$ if $\Psi_i^{(k)} \equiv \Psi_i^{(k_0)}$, for all $k \geq k_0$, k_0 fixed. Analogous to (3.70), r_i can be assessed in practice as

$$r_i = \lim_{k \rightarrow \infty} |\Psi_i^{(k+1)} - \Psi_i^{(k)}| / |\Psi_i^{(k)} - \Psi_i^{(k-1)}|.$$

Under broad regularity conditions, it can be shown (for example, Meng and Rubin, 1994) that

$$r = \max_{1 \leq i \leq d} r_i,$$

which is consistent with the intuition that the algorithm as a whole converges if and only if every component Ψ_i does. A component whose componentwise rate equals the global rate is then called the slowest component for the obvious reason. A component is the slowest if it is not orthogonal to the eigenvector corresponding to λ_{\max} , and thus typically there is more than one such component; see Meng (1994).

3.9.3 Rate Matrix in Terms of Information Matrices

Suppose that $\{\Psi^{(k)}\}$ is an EM sequence for which

$$\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi = \mathbf{0} \quad (3.71)$$

is satisfied by $\Psi = \Psi^{(k+1)}$, which will be the case with standard complete-data estimation. Then DLR showed that if $\Psi^{(k)}$ converges to a point Ψ^* , then

$$\mathbf{J}(\Psi^*) = \mathcal{I}_c^{-1}(\Psi^*; \mathbf{y}) \mathcal{I}_m(\Psi^*; \mathbf{y}). \quad (3.72)$$

This result was obtained also by Sundberg (1976).

Thus the rate of convergence of the EM algorithm is given by the largest eigenvalue of the information ratio matrix $\mathcal{I}_c^{-1}(\Psi; \mathbf{y}) \mathcal{I}_m(\Psi; \mathbf{y})$, which measures the proportion of information about Ψ that is missing by not also observing \mathbf{z} in addition to \mathbf{y} . The greater the proportion of missing information, the slower the rate of convergence. The fraction of information loss may vary across different components of Ψ , suggesting that certain components of Ψ may approach Ψ^* rapidly using the EM algorithm, while other components may require many iterations.

The rate of convergence of the EM algorithm can be expressed equivalently in terms of the largest eigenvalue of

$$\mathbf{I}_d - \mathcal{I}_c^{-1}(\Psi^*; \mathbf{y}) \mathbf{I}(\Psi^*; \mathbf{y}),$$

where \mathbf{I}_d denotes the $d \times d$ identity matrix. This is because we can express $\mathbf{J}(\Psi^*)$ also in the form

$$\mathbf{J}(\Psi^*) = \mathbf{I}_d - \mathcal{I}_c^{-1}(\Psi^*; \mathbf{y}) \mathbf{I}(\Psi^*; \mathbf{y}). \quad (3.73)$$

This result follows from (3.72) on noting from (3.50) that

$$\mathcal{I}_m(\Psi^*; \mathbf{y}) = \mathcal{I}_c(\Psi^*; \mathbf{y}) - \mathbf{I}(\Psi^*; \mathbf{y}).$$

From (3.73),

$$\begin{aligned} \partial^2 \log L(\Psi^*) / \partial \Psi \partial \Psi^T &= -\mathbf{I}(\Psi^*; \mathbf{y}) \\ &= -\mathcal{I}_c(\Psi^*; \mathbf{y}) \{ \mathbf{I}_d - \mathbf{J}(\Psi^*) \}. \end{aligned} \quad (3.74)$$

Exceptions to the convergence of the EM algorithm to a local maximum of $L(\Psi)$ occur if $J(\Psi^*)$ has eigenvalues exceeding unity. An eigenvalue of $J(\Psi^*)$ which is unity in a neighborhood of Ψ^* implies a ridge in $L(\Psi)$ through Ψ^* . Generally, we expect $\partial^2 \log L(\Psi^*)/\partial\Psi\partial\Psi^T$ to be negative semidefinite, if not negative definite, in which case the eigenvalues of $J(\Psi^*)$ all lie in $[0, 1]$ or $[0, 1)$, respectively.

The slow convergence of the EM algorithm has been reported in quite a few applications. Some of these are in Hartley (1958) (in contingency tables with missing data), Fienberg (1972) (in incomplete multiway contingency tables), Sundberg (1976), Haberman (1977), Nelder (1977) (when there is more than one missing observation in designed experiments; in latent structure analysis), and Thompson (1977) (in variance components estimation). Horng (1986, 1987) theoretically demonstrates that this behavior of the EM algorithm in some of these situations is due to the rate of convergence being 1 (called sublinear convergence). His examples include (a) mixtures of two normal distributions with a common variance in known proportions; (b) multivariate normal with missing values and with a singular covariance matrix but with nonsingular submatrices corresponding to the observed variables in each of the cases (which is what is required to execute the E-step); (c) a variance components model when one of the variances is close to zero at the limit point; and (d) some factor analysis situations.

3.9.4 Rate Matrix for Maximum *a Posteriori* Estimation

When finding the MAP estimate, the rate of convergence is given by replacing

$$\begin{aligned} -\mathcal{I}_c(\Psi^*; \mathbf{y}) &= -E_{\Psi^*}\{\mathbf{I}_c(\Psi^*; \mathbf{X}) | \mathbf{Y}\} \\ &= [\partial^2 Q(\Psi; \Psi^*)/\partial\Psi\partial\Psi^T]_{\Psi=\Psi^*} \end{aligned} \quad (3.75)$$

by

$$[\partial^2 Q(\Psi; \Psi^*)/\partial\Psi\partial\Psi^T]_{\Psi=\Psi^*} + \partial^2 \log p(\Psi^*)/\partial\Psi\partial\Psi^*$$

in (3.72), where $p(\Psi)$ is the prior density for Ψ .

3.9.5 Derivation of Rate Matrix in Terms of Information Matrices

We now show how the result (3.72), or equivalently (3.73), can be established. By a linear Taylor series expansion of the gradient vector $\mathbf{S}(\mathbf{y}; \Psi)$ of the log likelihood function $\log L(\Psi)$ about the point $\Psi = \Psi^{(k)}$, we have that

$$\mathbf{S}(\mathbf{y}; \Psi) \approx \mathbf{S}(\mathbf{y}; \Psi^{(k)}) - \mathbf{I}(\Psi^{(k)}; \mathbf{y})(\Psi - \Psi^{(k)}). \quad (3.76)$$

On putting $\Psi = \Psi^*$ in (3.76), it follows since $\mathbf{S}(\mathbf{y}; \Psi)$ vanishes at the point Ψ^* that

$$\Psi^* \approx \Psi^{(k)} + \mathbf{I}^{-1}(\Psi^{(k)}; \mathbf{y})\mathbf{S}(\mathbf{y}; \Psi^{(k)}). \quad (3.77)$$

The next step is to expand $\partial Q(\Psi; \Psi^{(k)})/\partial\Psi$ in a linear Taylor series about the point $\Psi = \Psi^{(k)}$ and to evaluate it at $\Psi = \Psi^{(k+1)}$ under the assumption (3.71) that it is zero at this point. This gives

$$\begin{aligned} \mathbf{0} &= [\partial Q(\Psi; \Psi^{(k)})/\partial\Psi]_{\Psi=\Psi^{(k+1)}} \\ &\approx [\partial Q(\Psi; \Psi^{(k)})/\partial\Psi]_{\Psi=\Psi^{(k)}} \\ &\quad + [\partial^2 Q(\Psi; \Psi^{(k)})/\partial\Psi\partial\Psi^T]_{\Psi=\Psi^{(k)}} (\Psi^{(k+1)} - \Psi^{(k)}) \\ &= \mathbf{S}(\mathbf{y}; \Psi^{(k)}) - \mathcal{I}_c(\Psi^{(k)}; \mathbf{y})(\Psi^{(k+1)} - \Psi^{(k)}), \end{aligned} \quad (3.78)$$

since from (3.75),

$$[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}} = -\mathcal{I}_c(\Psi^{(k)}; \mathbf{y}).$$

From (3.78)

$$S(\mathbf{y}; \Psi^{(k)}) \approx \mathcal{I}_c(\Psi^{(k)}; \mathbf{y})(\Psi^{(k+1)} - \Psi^{(k)}). \quad (3.79)$$

On using now this approximation for $S(\mathbf{y}; \Psi^{(k)})$ in (3.77), we have that

$$\begin{aligned} \Psi^* - \Psi^{(k)} &\approx \mathbf{I}^{-1}(\Psi^{(k)}; \mathbf{y})\mathcal{I}_c(\Psi^{(k)}; \mathbf{y})(\Psi^{(k+1)} - \Psi^{(k)}) \\ &= \mathbf{I}^{-1}(\Psi^{(k)}; \mathbf{y})\mathcal{I}_c(\Psi^{(k)}; \mathbf{y})(\Psi^{(k+1)} - \Psi^* + \Psi^* - \Psi^{(k)}). \end{aligned}$$

Thus

$$\begin{aligned} \Psi^{(k+1)} - \Psi^* &\approx \{\mathbf{I}_d - \mathcal{I}_c^{-1}(\Psi^{(k)}; \mathbf{y})\mathbf{I}(\Psi^{(k)}; \mathbf{y})\}(\Psi^{(k)} - \Psi^*) \\ &\approx \{\mathbf{I}_d - \mathcal{I}_c^{-1}(\Psi^*; \mathbf{y})\mathbf{I}(\Psi^*; \mathbf{y})\}(\Psi^{(k)} - \Psi^*) \\ &= \mathbf{J}(\Psi^*)(\Psi^{(k)} - \Psi^*), \end{aligned}$$

thus establishing (3.73).

3.9.6 Example 3.6: Censored Exponentially Distributed Survival Times (Example 1.3 Continued)

On putting $\mu = \hat{\mu}$ in the expression (3.55) for the conditional expectation of the (complete-data) information about μ , it can be seen that

$$I_c(\hat{\mu}; \mathbf{y}) = n\hat{\mu}^{-2},$$

while from (3.54), the incomplete-data observed information equals

$$I(\hat{\mu}; \mathbf{y}) = r\hat{\mu}^{-2}.$$

Corresponding to (3.67), we have

$$\mu^{(k+1)} - \hat{\mu} = J(\hat{\mu})(\mu^{(k)} - \hat{\mu}), \quad (3.80)$$

where

$$\begin{aligned} J(\hat{\mu}) &= 1 - I_c^{-1}(\hat{\mu}; \mathbf{y})I(\hat{\mu}; \mathbf{y}) \\ &= 1 - n^{-1}r. \end{aligned}$$

The result (3.80) was established directly in Section 1.5; see equation (1.53).

This Page Intentionally Left Blank

STANDARD ERRORS AND SPEEDING UP CONVERGENCE

4.1 INTRODUCTION

In this chapter, we address two issues that have led to some criticism of the EM algorithm. The first concerns the provision of standard errors, or the full covariance matrix in multivariate situations, of the MLE obtained via the EM algorithm. One initial criticism of the EM algorithm was that it does not automatically provide an estimate of the covariance matrix of the MLE, as do some other methods, such as Newton-type methods.

Hence we shall consider a number of methods for assessing the covariance matrix of the MLE $\hat{\Psi}$ of the parameter vector Ψ , obtained via the EM algorithm. Most of these methods are based on the observed information matrix $I(\hat{\Psi}; \mathbf{y})$. It will be seen that there are methods that allow $I(\hat{\Psi}; \mathbf{y})$ to be calculated within the EM framework. For independent data, $I(\hat{\Psi}; \mathbf{y})$ can be approximated without additional work beyond the calculations used to compute the MLE in the first instance.

The other common criticism that has been leveled at the EM algorithm is that its convergence can be quite slow. We shall therefore consider some methods that have been proposed for accelerating the EM algorithm. However, methods to accelerate the EM algorithm do tend to sacrifice the simplicity it usually enjoys. As remarked by Lange (1995b), it is likely that no acceleration method can match the stability and simplicity of the unadorned EM algorithm.

The methods to be discussed in this chapter for speeding up the convergence of the EM algorithm are applicable for a given specification of the complete data. In the next chapter, we consider some developments that approach the problem of speeding up convergence in

terms of the choice of the missing data in the specification of the complete-data problem in the EM framework. They introduce a working parameter in the specification of the complete data, which thus indexes a class of EM algorithms. We shall see that in some particular cases it is possible to modify the choice of the complete data so that the resulting EM algorithm can be just as easily implemented yet produce striking gains in the speed of convergence.

4.2 OBSERVED INFORMATION MATRIX

4.2.1 Direct Evaluation

As explained in Section 4.1, it is common in practice to estimate the inverse of the covariance matrix of the MLE $\hat{\Psi}$ by the observed information matrix $I(\hat{\Psi}; \mathbf{y})$. Hence we shall consider in the following subsections a number of ways for calculating or approximating $I(\hat{\Psi}; \mathbf{y})$. One way to proceed is to directly evaluate $I(\hat{\Psi}; \mathbf{y})$ after the computation of the MLE $\hat{\Psi}$. However, analytical evaluation of the second-order derivatives of the incomplete-data log likelihood, $\log L(\Psi)$, may be difficult, or at least tedious. Indeed, often it is for reasons of this nature that the EM algorithm is used to compute the MLE in the first instance.

4.2.2 Extraction of Observed Information Matrix in Terms of the Complete-Data Log Likelihood

Louis (1982) shows that the missing information matrix $\mathcal{I}_m(\Psi; \mathbf{y})$ can be expressed in the form

$$\mathcal{I}_m(\Psi; \mathbf{y}) = \text{cov}_{\Psi}\{S_c(\mathbf{X}; \Psi) | \mathbf{y}\} \quad (4.1)$$

$$= E_{\Psi}\{S_c(\mathbf{X}; \Psi)S_c^T(\mathbf{X}; \Psi) | \mathbf{y}\} \\ - S(\mathbf{y}; \Psi)S^T(\mathbf{y}; \Psi), \quad (4.2)$$

since

$$S(\mathbf{y}; \Psi) = E_{\Psi}\{S_c(\mathbf{X}; \Psi) | \mathbf{y}\}.$$

On substituting (4.1) and then (4.2) into (3.50), we have that

$$\begin{aligned} I(\Psi; \mathbf{y}) &= \mathcal{I}_c(\Psi; \mathbf{y}) - \mathcal{I}_m(\Psi; \mathbf{y}) \\ &= \mathcal{I}_c(\Psi; \mathbf{y}) - \text{cov}_{\Psi}\{S_c(\mathbf{X}; \Psi) | \mathbf{y}\} \\ &= \mathcal{I}_c(\Psi; \mathbf{y}) \\ &\quad - E_{\Psi}\{S_c(\mathbf{X}; \Psi)S_c^T(\mathbf{X}; \Psi) | \mathbf{y}\} \\ &\quad + S(\mathbf{y}; \Psi)S^T(\mathbf{y}; \Psi). \end{aligned} \quad (4.3)$$

The result (4.3) can be established by directly showing that $\mathcal{I}_m(\Psi; \mathbf{y})$ can be put in the form of the right-hand side of (4.1). Louis (1982) establishes (4.3) by working with

$\mathbf{I}(\Psi; \mathbf{y})$ as follows. From (3.45), $\mathbf{I}(\Psi; \mathbf{y})$ can be expressed as

$$\begin{aligned}
\mathbf{I}(\Psi; \mathbf{y}) &= -\partial S(\mathbf{y}; \Psi)/\partial\Psi \\
&= -\partial[\{\int_{\mathcal{X}(\mathbf{y})} g'_c(\mathbf{x}; \Psi) d\mathbf{x}\}/g(\mathbf{y}; \Psi)]\partial\Psi \\
&= -\{\int_{\mathcal{X}(\mathbf{y})} \partial^2 g_c(\mathbf{x}; \Psi)/\partial\Psi\partial\Psi^T d\mathbf{x}\}/g(\mathbf{y}; \Psi) \\
&\quad + \{\int_{\mathcal{X}(\mathbf{y})} g'_c(\mathbf{x}; \Psi) d\mathbf{x}\} \{\int_{\mathcal{X}(\mathbf{y})} g'_c(\mathbf{x}; \Psi) d\mathbf{x}\}^T / \{g(\mathbf{y}; \Psi)\}^2.
\end{aligned} \tag{4.4}$$

In (4.4) and the equations below, the prime denotes differentiation with respect to Ψ . Also, it is assumed in this section and the sequel that regularity conditions hold for the interchange of the operations of differentiation and integration where necessary.

Proceeding as before with the derivation of (3.46),

$$\mathbf{I}(\Psi; \mathbf{y}) = -\{\int_{\mathcal{X}(\mathbf{y})} \partial^2 g_c(\mathbf{x}; \Psi)/\partial\Psi\partial\Psi^T d\mathbf{x}\}/g(\mathbf{y}; \Psi) + \mathbf{S}(\mathbf{y}; \Psi)\mathbf{S}^T(\mathbf{y}; \Psi), \tag{4.5}$$

on using the result (3.45) for the last term on the right-hand side of (4.4).

The first term on the right-hand side of (4.5) can be expressed as

$$\begin{aligned}
&-\{\int_{\mathcal{X}(\mathbf{y})} \partial^2 g_c(\mathbf{x}; \Psi)/\partial\Psi\partial\Psi^T d\mathbf{x}\}/g(\mathbf{y}; \Psi) \\
&= -\int_{\mathcal{X}(\mathbf{y})} [\{\partial^2 \log g_c(\mathbf{x}; \Psi)/\partial\Psi\partial\Psi^T\} \{g_c(\mathbf{x}; \Psi)/g(\mathbf{y}; \Psi)\}] d\mathbf{x} \\
&\quad - \int_{\mathcal{X}(\mathbf{y})} \{g'_c(\mathbf{x}; \Psi)/g_c(\mathbf{x}; \Psi)\} \{g'_c(\mathbf{x}; \Psi)/g_c(\mathbf{x}; \Psi)\}^T \{g_c(\mathbf{x}; \Psi)/g(\mathbf{y}; \Psi)\} d\mathbf{x} \\
&= \int_{\mathcal{X}(\mathbf{y})} \mathbf{I}_c(\Psi; \mathbf{x}) k(\mathbf{x} | \mathbf{y}; \Psi) d\mathbf{x} \\
&\quad - \int_{\mathcal{X}(\mathbf{y})} \mathbf{S}_c(\mathbf{x}; \Psi) \mathbf{S}_c^T(\mathbf{x}; \Psi) k(\mathbf{x} | \mathbf{y}; \Psi) d\mathbf{x} \\
&= E_{\Psi} \{\mathbf{I}_c(\Psi; \mathbf{X}) | \mathbf{y}\} - E_{\Psi} \{\mathbf{S}_c(\mathbf{X}; \Psi) \mathbf{S}_c^T(\mathbf{X}; \Psi) | \mathbf{y}\}. \\
&= \mathcal{I}_c(\Psi; \mathbf{y}) - E_{\Psi} \{\mathbf{S}_c(\mathbf{X}; \Psi) \mathbf{S}_c^T(\mathbf{X}; \Psi) | \mathbf{y}\}.
\end{aligned} \tag{4.6}$$

Substitution of (4.6) into (4.5) gives the expression (4.3) for $\mathbf{I}(\Psi; \mathbf{y})$.

From (4.3), the observed information matrix $\mathbf{I}(\hat{\Psi})$ can be computed as

$$\mathbf{I}(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c(\hat{\Psi}; \mathbf{y}) - \mathcal{I}_m(\hat{\Psi}; \mathbf{y}) \tag{4.7}$$

$$= \mathcal{I}_c(\hat{\Psi}; \mathbf{y}) - [\text{cov}_{\Psi} \{\mathbf{S}_c(\mathbf{X}; \Psi) | \mathbf{y}\}]_{\Psi=\hat{\Psi}} \tag{4.8}$$

$$= \mathcal{I}_c(\hat{\Psi}; \mathbf{y}) - [E_{\Psi} \{\mathbf{S}_c(\mathbf{X}; \Psi) \mathbf{S}_c^T(\mathbf{X}; \Psi)\} | \mathbf{y}]_{\Psi=\hat{\Psi}}, \tag{4.9}$$

since the last term on the right-hand side of (4.3) is zero as $\hat{\Psi}$ satisfies

$$\mathbf{S}(\mathbf{y}; \Psi) = \mathbf{0}.$$

Hence the observed information matrix for the original (incomplete-data) problem can be computed in terms of the conditional moments of the gradient and curvature of the complete-data log likelihood function introduced within the EM framework.

It is important to emphasize that the aforementioned estimates of the covariance matrix of the MLE are based on the second derivatives of the (complete-data) log likelihood, and so are guaranteed to be valid inferentially only asymptotically. Consequently, from both frequentist and Bayesian perspectives, the practical propriety of the resulting normal theory inferences is improved when the log likelihood is more nearly normal. To this end, Meng and Rubin (1991) transform some of the parameters in their examples, which are to be presented in Section 4.5.

4.2.3 Regular Case

In the case of the regular exponential family with Ψ as the natural parameter vector, the matrix $I_c(\Psi; \mathbf{x})$ is not a function of the data and so

$$\mathcal{I}_c(\Psi; \mathbf{y}) = \mathcal{I}_c(\Psi) \quad (4.10)$$

$$\begin{aligned} &= \text{cov}_{\Psi}\{S_c(\mathbf{X}; \Psi)\} \\ &= \text{cov}_{\Psi}\{t(\mathbf{X})\} \end{aligned} \quad (4.11)$$

since from (1.56),

$$S_c(\mathbf{X}; \Psi) = t(\mathbf{X}).$$

From (4.1), we have that

$$\begin{aligned} \mathcal{I}_m(\Psi; \mathbf{y}) &= \text{cov}_{\Psi}\{S_c(\mathbf{X}; \Psi) | \mathbf{y}\} \\ &= \text{cov}_{\Psi}\{t(\mathbf{X}) | \mathbf{y}\}. \end{aligned} \quad (4.12)$$

Hence from (4.7), the observed incomplete-data information matrix $I(\hat{\Psi}; \mathbf{y})$ can be written as the difference at $\Psi = \hat{\Psi}$ between the unconditional covariance matrix and the conditional covariance matrix of the complete-data score function $S_c(\mathbf{X}; \Psi)$, which is the sufficient statistic $t(\mathbf{X})$. Thus the observed information matrix $I(\hat{\Psi}; \mathbf{y})$ is given by

$$I(\hat{\Psi}; \mathbf{y}) = [\text{cov}_{\Psi}\{t(\mathbf{X})\} - \text{cov}_{\Psi}\{t(\mathbf{X}) | \mathbf{y}\}]_{\Psi=\hat{\Psi}}.$$

4.2.4 Evaluation of the Conditional Expected Complete-Data Information Matrix

If any of the formulas (4.7) to (4.9) is to be used to calculate the observed information matrix $I(\hat{\Psi}; \mathbf{y})$, then we have to calculate $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$, the conditional expectation of the complete-data information matrix evaluated at the MLE. Also, it will be seen shortly in Section 4.5 that we need to calculate $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$ when using the SupplementedEM algorithm to compute the observed information matrix $I(\hat{\Psi}; \mathbf{y})$.

The calculation of $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$ is readily facilitated by standard complete-data computations if the complete-data density $g_c(\mathbf{x}; \Psi)$ belongs to the regular exponential family, since then

$$\mathcal{I}_c(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c(\hat{\Psi}). \quad (4.13)$$

This result is obvious from (4.10) if Ψ is the natural parameter. Suppose now that $\theta = c(\Psi)$ is the natural parameter. Then to show that (4.13) still holds, we have on using

the chain rule to differentiate minus $\log L_c(\Psi)$ twice with respect to Ψ that

$$\mathbf{I}_c(\Psi; \mathbf{x}) = \mathbf{A}^T (-\partial^2 \log L_c(\Psi) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T) \mathbf{A} + \mathbf{B}, \quad (4.14)$$

where

$$\begin{aligned} (\mathbf{A})_{ij} &= \partial \theta_i / \partial \Psi_j, \\ (\mathbf{B})_{ij} &= \mathbf{d}_{ij}^T \partial \log L_c(\Psi) / \partial \boldsymbol{\theta}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{d}_{ij} &= (\partial^2 \theta_1 / \partial \Psi_i \partial \Psi_j, \dots, \partial^2 \theta_d / \partial \Psi_i \partial \Psi_j)^T \\ &= \partial^2 \boldsymbol{\theta} / \partial \Psi_i \partial \Psi_j \end{aligned}$$

for $i, j = 1, \dots, d$. As $-\partial^2 \log L_c(\Psi) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ does not depend on the data, it equals the expected complete-data information matrix for $\boldsymbol{\theta}$, and so

$$\mathbf{A}^T (-\partial^2 \log L_c(\Psi) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T) \mathbf{A}$$

is equal to $\mathcal{I}_c(\Psi)$, the expected complete-data information matrix for Ψ . It therefore follows on taking the conditional expectation of (4.14) given \mathbf{y} that

$$\mathcal{I}_c(\Psi; \mathbf{y}) = \mathcal{I}_c(\Psi) + E_{\Psi}(\mathbf{B} | \mathbf{y}), \quad (4.15)$$

where

$$\begin{aligned} E_{\Psi}(\mathbf{B} | \mathbf{y}) &= \mathbf{d}_{ij}^T E_{\Psi} \{ \partial \log L_c(\Psi) / \partial \boldsymbol{\theta} | \mathbf{y} \} \\ &= \mathbf{d}_{ij}^T \partial \log L(\Psi) / \partial \boldsymbol{\theta}. \end{aligned} \quad (4.16)$$

Now

$$\partial \log L(\Psi) / \partial \boldsymbol{\theta} = \mathbf{0}$$

at $\boldsymbol{\theta}^* = c(\Psi^*)$ for any stationary point Ψ^* of $L(\Psi)$. Thus then from (4.15) and (4.16),

$$\mathcal{I}_c(\Psi^*; \mathbf{y}) = \mathcal{I}_c(\Psi^*),$$

for any stationary point Ψ^* of $L(\Psi)$, including the MLE $\hat{\Psi}$ of Ψ . When $g_c(\mathbf{x}; \Psi)$ is from an irregular exponential family, the calculation of $\mathcal{I}_c(\Psi; \mathbf{y})$ is still straightforward, as $\log L_c(\Psi)$ is linear in $\mathbf{t}(\mathbf{x})$. Hence $\mathcal{I}_c(\Psi; \mathbf{y})$ is formed from $\mathbf{I}_c(\Psi; \mathbf{x})$ simply by replacing $\mathbf{t}(\mathbf{x})$ by its conditional expectation.

4.2.5 Examples

Example 4.1: Information Matrix for the Multinomial Example (*Example 1.1 Continued*). For this example, the observed information $I(\hat{\Psi}; \mathbf{y})$ can be obtained directly by evaluating the right-hand side of (1.16) at $\Psi = \hat{\Psi} = 0.627$ to give

$$I(\hat{\Psi}; \mathbf{y}) = 377.5. \quad (4.17)$$

Thus the asymptotic standard error of $\hat{\Psi}$ is

$$1/\sqrt{377.5} = 0.051. \quad (4.18)$$

Concerning the expression (4.8) for the asymptotic variance of $\hat{\Psi}$, we now consider the calculation of $I_c(\hat{\Psi}; \mathbf{y})$. On differentiation of (1.22) with respect to Ψ ,

$$S_c(\mathbf{x}; \Psi) = \frac{y_{12} + y_4}{\Psi} - \frac{y_2 + y_3}{1 - \Psi} \quad (4.19)$$

and

$$I_c(\Psi; \mathbf{x}) = \frac{y_{12} + y_4}{\Psi^2} + \frac{y_2 + y_3}{(1 - \Psi)^2}. \quad (4.20)$$

Taking the conditional expectation of (4.20) given \mathbf{y} gives

$$\mathcal{I}_c(\Psi; \mathbf{y}) = \frac{E_\Psi(Y_{12} | \mathbf{y}) + y_4}{\Psi^2} + \frac{y_2 + y_3}{(1 - \Psi)^2}, \quad (4.21)$$

where

$$E_\Psi(Y_{12} | \mathbf{y}) = \frac{1}{4}y_1\Psi / (\frac{1}{2} + \frac{1}{4}\Psi). \quad (4.22)$$

This last result (4.22) follows from the work in Section 1.4.2 where it was noted that the conditional distribution of Y_{12} given \mathbf{y} is binomial with sample size $y_1 = 125$ and probability parameter

$$\frac{1}{4}\Psi / (\frac{1}{2} + \frac{1}{4}\Psi).$$

On evaluation at $\Psi = \hat{\Psi} = 0.6268214980$, we obtain

$$I_c(\hat{\Psi}; \mathbf{y}) = 435.318. \quad (4.23)$$

From (4.21), we can express $I_c(\hat{\Psi}; \mathbf{y})$ also as

$$I_c(\hat{\Psi}; \mathbf{y}) = \frac{E_{\hat{\Psi}}(Y_{12} | \mathbf{y}) + y_2 + y_3 + y_4}{\hat{\Psi}(1 - \hat{\Psi})}. \quad (4.24)$$

This agrees with the fact that $L_c(\Psi)$ has a binomial form with probability parameter Ψ and sample size $Y_{12} + y_2 + y_3 + y_4$, with conditional expectation

$$E_\Psi(Y_{12} | \mathbf{y}) + y_2 + y_3 + y_4.$$

Thus $I_c(\Psi; \mathbf{y})$ is given by the conditional expectation of the inverse of the binomial variance,

$$(Y_{12} + y_2 + y_3 + y_4) / \{\Psi(1 - \Psi)\}.$$

It is also of interest to note here that

$$I_c(\hat{\Psi}; \mathbf{y}) \neq I_c(\hat{\Psi}).$$

For on taking the expectation of (4.20), we have that

$$I_c(\Psi) = \frac{1}{2}n / \{\Psi(1 - \Psi)\},$$

and so

$$I_c(\hat{\Psi}) = \frac{1}{2}n / \{\hat{\Psi}(1 - \hat{\Psi})\},$$

which does not equal (4.24). This can be explained by the fact that the complete-data density $g_c(\mathbf{x}; \Psi)$ is a member of an irregular exponential family.

Moving now to the computation of $I_m(\hat{\Psi}; \mathbf{y})$, we have from (4.1) and (4.19) that

$$\begin{aligned} I_m(\Psi; \mathbf{y}) &= \text{var}_\Psi \{S_c(\mathbf{X}; \Psi) | \mathbf{y}\} \\ &= \text{var}_\Psi \left\{ \left(\frac{Y_{12} + y_4}{\Psi} - \frac{y_2 + y_3}{1 - \Psi} \right) | \mathbf{y} \right\} \\ &= \Psi^{-2} \text{var}_\Psi(Y_{12} | \mathbf{y}) \\ &= \Psi^{-2} y_1 \{(\Psi/8)/(\frac{1}{2} + \frac{1}{4}\Psi)^2\}, \end{aligned} \quad (4.25)$$

which on evaluation at $\Psi = \hat{\Psi}$ gives

$$I_m(\hat{\Psi}; \mathbf{y}) = 57.801. \quad (4.26)$$

From (4.23) and (4.26), we obtain that the observed information is given by

$$\begin{aligned} I(\hat{\Psi}; \mathbf{y}) &= I_c(\hat{\Psi}; \mathbf{y}) - I_m(\hat{\Psi}; \mathbf{y}) \\ &= 435.318 - 57.801 \\ &= 377.517, \end{aligned}$$

which yields 0.051 as the standard error of $\hat{\Psi}$. This agrees with (4.18) obtained by direct evaluation of the observed information $I(\hat{\Psi}; \mathbf{y})$.

From (3.72), we have that the rate of convergence is given approximately by

$$\begin{aligned} J(\hat{\Psi}) &= I_m(\hat{\Psi})/I_c(\hat{\Psi}) \\ &= 57.801/435.318 \\ &= 0.1328, \end{aligned}$$

which is in agreement with the assessed rate of 0.1328 in Table 1.2.

Example 4.2. Mixture of Two Univariate Normals with Known Common Variance. One of the examples of Louis (1982) to demonstrate the use of the formula (4.8) or (4.9) for calculating the observed information matrix $\mathbf{I}(\hat{\Psi}; \mathbf{y})$ was ML estimation for a mixture of two univariate normal densities with known common variance set equal to one. That is, the observed data vector is given by

$$\mathbf{y} = (w_1, \dots, w_n)^T,$$

where w_1, \dots, w_n denote the observed values of a random sample from the density

$$f(w; \boldsymbol{\Psi}) = \pi_1 \phi(w; \mu_1, 1) + \pi_2 \phi(w; \mu_2, 1),$$

where

$$\boldsymbol{\Psi} = (\pi_1, \mu_1, \mu_2)^T.$$

In illustrating the use of the result (4.8) for this particular example, we shall just consider the diagonal elements of $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$, as it is easy to check that its off-diagonal elements are all zero.

Concerning the calculation of $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$ first, we have that the diagonal elements of $\mathcal{I}_c(\boldsymbol{\Psi}; \mathbf{x})$ are given by

$$I_{c;11}(\boldsymbol{\Psi}; \mathbf{x}) = \sum_{j=1}^n (z_{1j}/\pi_1^2 + z_{2j}/\pi_2^2), \quad (4.27)$$

$$I_{c;22}(\boldsymbol{\Psi}; \mathbf{x}) = \sum_{j=1}^n z_{1j}, \quad (4.28)$$

and

$$I_{c;33}(\Psi; \mathbf{x}) = \sum_{j=1}^n z_{2j}. \quad (4.29)$$

In the above equations, $I_{c;ii}(\Psi; \mathbf{x})$ denotes the i th diagonal element of $\mathbf{I}_c(\Psi; \mathbf{x})$.

On taking the conditional expectation of (4.27) to (4.29) given \mathbf{y} , we find that

$$I_{c;11}(\Psi; \mathbf{y}) = \sum_{j=1}^n \{\tau_1(w_j; \Psi)/\pi_1^2 + \tau_2(w_j; \Psi)/\pi_2^2\} \quad (4.30)$$

$$I_{c;22}(\Psi; \mathbf{y}) = \sum_{j=1}^n \tau_1(w_j; \Psi), \quad (4.31)$$

and

$$I_{c;33}(\Psi; \mathbf{y}) = \sum_{j=1}^n \tau_2(w_j; \Psi). \quad (4.32)$$

where $\tau_1(w_j, \Psi)$ and $\tau_2(w_j, \Psi)$ are as in Section 2.7. Evaluation of the right-hand sides of (4.30) to (4.32) at the point $\Psi = \hat{\Psi}$ gives

$$\mathcal{I}_c(\hat{\Psi}; \mathbf{y}) = \text{diag}\{n/(\hat{\pi}_1\hat{\pi}_2), n\hat{\pi}_1, n\hat{\pi}_2\}. \quad (4.33)$$

In this particular example where the complete-data density belongs to the regular exponential family, it is simpler to use the relationship

$$\mathcal{I}_c(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c(\hat{\Psi})$$

to calculate $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$, since it is well known that

$$\mathcal{I}_c(\Psi) = \text{diag}\{n/(\pi_1\pi_2), n\pi_1, n\pi_2\}.$$

This last result can be verified by taking the unconditional expectation of (4.27) to (4.29).

Considering the second term on the right-hand side of (4.8), we have that the elements of the complete-data gradient vector $\mathbf{S}_c(\mathbf{x}; \Psi)$ are given by

$$\begin{aligned} S_{c;1}(\mathbf{x}; \Psi) &= \sum_{j=1}^n (z_{1j}/\pi_1 - z_{2j}/\pi_2) \\ &= (\sum_{j=1}^n z_{1j} - n\pi_1)/(\pi_1\pi_2) \end{aligned}$$

and

$$S_{c;i+1}(\mathbf{x}; \Psi) = \sum_{j=1}^n (w_j - \mu_{i+1})z_{ij} \quad (i = 1, 2).$$

We now proceed to the calculation of $\mathcal{I}_m(\hat{\Psi}; \mathbf{y})$ via the formula

$$\mathcal{I}_m(\Psi; \mathbf{y}) = \text{cov}_{\Psi}\{\mathbf{S}_c(\mathbf{X}; \Psi) \mid \mathbf{y}\}.$$

The first diagonal element of $\mathcal{I}_m(\Psi; \mathbf{y})$ is given by

$$\begin{aligned} I_{m;11}(\Psi; \mathbf{y}) &= \text{var}_{\Psi}\left\{\left(\sum_{j=1}^n Z_{1j} - n\pi_1\right)/(\pi_1\pi_2) \mid \mathbf{y}\right\} \\ &= \pi_1^{-2}\pi_2^2 \text{var}_{\Psi}\left\{\sum_{j=1}^n Z_{ij} \mid \mathbf{y}\right\} \\ &= \pi_1^{-2}\pi_2^{-2} \sum_{j=1}^n \tau_1(w_j; \Psi)\tau_2(w_j; \Psi). \end{aligned}$$

Proceeding in a similar manner, it can be shown that

$$\begin{aligned} I_{m;i+1,i+1}(\Psi; \mathbf{y}) &= \text{var}_{\Psi}\left\{\sum_{j=1}^n (w_j - \mu_i)Z_{ij} \mid \mathbf{y}\right\} \\ &= \sum_{j=1}^n \tau_1(w_j; \Psi)\tau_2(w_j; \Psi)(w_j - \mu_i)^2 \end{aligned} \quad (4.34)$$

$$\begin{aligned} I_{m;1,i+1}(\Psi; \mathbf{y}) &= \text{cov}_{\Psi}\left\{\sum_{j=1}^n \frac{Z_{1j} - n\pi_1}{\pi_1\pi_2}, \sum_{j=1}^n (w_j - \mu_i)Z_{ij}\right\} \\ &= (-1)^{i+1}\pi_1^{-1}\pi_2^{-1} \sum_{j=1}^n \tau_1(w_j; \Psi)\tau_2(w_j; \Psi)(w_j - \mu_i) \end{aligned} \quad (4.35)$$

for $i = 1, 2$ and

$$\begin{aligned} I_{m;23}(\Psi; \mathbf{y}) &= \text{cov}_{\Psi}\left\{\sum_{j=1}^n (w_j - \mu_1)Z_{1j}, \sum_{j=1}^n (w_j - \mu_2)Z_{2j}\right\} \\ &= -\sum_{j=1}^n \tau_1(w_j; \Psi)\tau_2(w_j; \Psi)(w_j - \mu_1)(w_j - \mu_2). \end{aligned}$$

The observed information matrix $\mathbf{I}(\hat{\Psi}; \mathbf{y})$ can be obtained on evaluating $\mathcal{I}_m(\Psi; \mathbf{y})$ at $\Psi = \hat{\Psi}$ and subtracting it from $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$.

Of course in this simplified example, $\mathbf{I}(\hat{\Psi}; \mathbf{y})$ could have been obtained by direct differentiation of the (incomplete-data) log likelihood. To illustrate this for the second diagonal term, we have that

$$\begin{aligned} I_{22}(\Psi; \mathbf{y}) &= -\partial^2 \log L(\Psi)/\partial \mu_1^2 \\ &= n\pi_1 - \sum_{j=1}^n \tau_1(w_j; \Psi)\tau_2(w_j; \Psi)(w_j - \mu_1)^2. \end{aligned} \quad (4.36)$$

From (4.33) and (4.34), it can be seen that

$$I_{22}(\hat{\Psi}; \mathbf{y}) = I_{c;22}(\hat{\Psi}; \mathbf{y}) - I_{m;22}(\hat{\Psi}; \mathbf{y}).$$

4.3 APPROXIMATIONS TO THE OBSERVED INFORMATION MATRIX: I.I.D. CASE

It can be seen from the expression (4.9) for the observed information matrix $I(\hat{\Psi}; \mathbf{y})$ that it requires in addition to the code for the E- and M- steps, the calculation of the conditional (on the observed data \mathbf{y}) expectation of the complete-data information matrix $I(\Psi; \mathbf{x})$ and of the complete-data score statistic vector $S_c(\mathbf{X}; \Psi)$ times its transpose. For many problems, this can be algebraically tedious or even intractable. Hence we now consider some practical methods for approximating the observed information matrix.

In the case of independent and identically distributed (i.i.d.) data, an approximation to the observed information matrix is readily available without any additional analyses having to be performed. In the i.i.d. case, the observed data \mathbf{y} may be assumed to consist of observations $\mathbf{w}_1, \dots, \mathbf{w}_n$ on n independent and identically distributed random variables with common p.d.f., say, $f(\mathbf{w}; \Psi)$. The log likelihood $\log L(\Psi)$ can be expressed then in the form

$$\log L(\Psi) = \sum_{j=1}^n \log L_j(\Psi),$$

where

$$L_j(\Psi) = f(\mathbf{w}_j; \Psi)$$

is the likelihood function for Ψ formed from the single observation \mathbf{w}_j ($j = 1, \dots, n$).

We can now write the score vector $\mathbf{S}(\mathbf{y}; \Psi)$ as

$$\mathbf{S}(\mathbf{y}; \Psi) = \sum_{j=1}^n \mathbf{s}(\mathbf{w}_j; \Psi),$$

where

$$\mathbf{s}(\mathbf{w}_j; \Psi) = \partial \log L_j(\Psi) / \partial \Psi.$$

The expected information matrix $\mathcal{I}(\Psi)$ can be written as

$$\mathcal{I}(\Psi) = ni(\Psi), \quad (4.37)$$

where

$$\begin{aligned} i(\Psi) &= E_{\Psi} \{ \mathbf{s}(\mathbf{W}; \Psi) \mathbf{s}^T(\mathbf{W}; \Psi) \} \\ &= \text{cov}_{\Psi} \{ \mathbf{s}(\mathbf{W}; \Psi) \} \end{aligned} \quad (4.38)$$

is the information contained in a single observation. Corresponding to (4.38), the empirical information matrix (in a single observation) can be defined to be

$$\begin{aligned} \bar{i}(\Psi) &= n^{-1} \sum_{j=1}^n \mathbf{s}(\mathbf{w}_j; \Psi) \mathbf{s}^T(\mathbf{w}_j; \Psi) - \bar{\mathbf{s}} \bar{\mathbf{s}}^T \\ &= n^{-1} \sum_{j=1}^n \mathbf{s}(\mathbf{w}_j; \Psi) \mathbf{s}^T(\mathbf{w}_j; \Psi) \\ &\quad - n^{-2} \mathbf{S}(\mathbf{y}; \Psi) \mathbf{S}^T(\mathbf{y}; \Psi), \end{aligned} \quad (4.39)$$

where

$$\bar{\mathbf{s}} = n^{-1} \sum_{j=1}^n \mathbf{s}(\mathbf{w}_j; \Psi).$$

Corresponding to this empirical form (4.39) for $\mathbf{i}(\Psi)$, $\mathcal{I}(\Psi)$ is estimated by

$$\begin{aligned}\mathbf{I}_e(\Psi; \mathbf{y}) &= n\bar{\mathbf{i}}(\Psi) \\ &= \sum_{j=1}^n \mathbf{s}(\mathbf{w}_j; \Psi) \mathbf{s}^T(\mathbf{w}_j; \Psi) \\ &\quad - n^{-1} \mathbf{S}(\mathbf{y}; \Psi) \mathbf{S}^T(\mathbf{y}; \Psi).\end{aligned}\quad (4.40)$$

On evaluation at $\Psi = \hat{\Psi}$, $\mathbf{I}_e(\hat{\Psi}; \mathbf{y})$ reduces to

$$\mathbf{I}_e(\hat{\Psi}; \mathbf{y}) = \sum_{j=1}^n \mathbf{s}(\mathbf{w}_j; \hat{\Psi}) \mathbf{s}^T(\mathbf{w}_j; \hat{\Psi}), \quad (4.41)$$

since $\mathbf{S}(\mathbf{y}; \hat{\Psi}) = \mathbf{0}$.

Meilijson (1989) terms $\mathbf{I}_e(\hat{\Psi}; \mathbf{y})$ as the empirical observed information matrix. It is used commonly in practice to approximate the observed information matrix $\mathbf{I}(\hat{\Psi}; \mathbf{y})$; see, for example, Berndt, Hall, Hall, and Hausman (1974), Redner and Walker (1984), and McLachlan and Basford (1988, Chapter 2).

It follows that $\mathbf{I}_e(\hat{\Psi}; \mathbf{y})/n$ is a consistent estimator of $\mathbf{i}(\Psi)$. The use of (4.41) can be justified also in the following sense. Now

$$\begin{aligned}\mathbf{I}(\Psi; \mathbf{y}) &= -\partial^2 \log L(\Psi)/\partial\Psi\partial\Psi^T \\ &= -\sum_{j=1}^n \partial^2 \log L_j(\Psi)/\partial\Psi\partial\Psi^T \\ &= \sum_{j=1}^n \{\partial \log L_j(\Psi)/\partial\Psi\} \{\partial \log L_j(\Psi)/\partial\Psi\}^T \\ &\quad - \sum_{j=1}^n \{\partial^2 L_j(\Psi)/\partial\Psi\partial\Psi^T\}/L_j(\Psi).\end{aligned}\quad (4.42)$$

The second term on the right-hand side of (4.42) has zero expectation. Hence

$$\begin{aligned}\mathbf{I}(\hat{\Psi}; \mathbf{y}) &\approx \sum_{j=1}^n \{\partial \log L_j(\hat{\Psi})/\partial\Psi\} \{\partial \log L_j(\hat{\Psi})/\partial\Psi\}^T \\ &= \sum_{j=1}^n \mathbf{s}(\mathbf{w}_j; \hat{\Psi}) \mathbf{s}^T(\mathbf{w}_j; \hat{\Psi}) \\ &= \mathbf{I}_e(\hat{\Psi}; \mathbf{y}),\end{aligned}\quad (4.43)$$

where the accuracy of this approximation depends on how close $\hat{\Psi}$ is to Ψ .

In the particular case where the complete-data density $g_c(\mathbf{x}; \Psi)$ is of multinomial form, the second term on the right-hand side of (4.42) is zero, and so (4.43) holds exactly.

It follows from the result (3.44) that

$$s_j(\mathbf{w}_j; \Psi) = E_\Psi \{\partial \log L_{cj}(\Psi)/\partial\Psi \mid \mathbf{y}\},$$

where $L_{cj}(\Psi)$ is the complete-data likelihood formed from the single observation \mathbf{w}_j ($j = 1, \dots, n$). Thus the approximation $\mathbf{I}_e(\hat{\Psi}; \mathbf{y})$ to the observed information matrix $\mathbf{I}(\hat{\Psi}; \mathbf{y})$ can be expressed in terms of the conditional expectation of the gradient vector of the complete-data log likelihood function evaluated at the MLE $\hat{\Psi}$. It thus avoids the computation of second-order partial derivatives of the complete-data log likelihood.

4.4 OBSERVED INFORMATION MATRIX FOR GROUPED DATA

4.4.1 Approximation Based on Empirical Information

We consider here an approximation to the observed information matrix in the case where the available data are truncated and grouped into r intervals, as formulated in Section 2.8. We let

$$\begin{aligned}s_j(\Psi^{(k)}) &= [\partial Q_j(\Psi; \Psi^{(k)})/\partial \Psi]_{\Psi=\Psi^{(k)}} \\ &= \partial \log P_j(\Psi)/\partial \Psi\end{aligned}\quad (4.44)$$

from (2.73), where $Q_j(\Psi; \Psi^{(k)})$ is defined by (2.69). It can be seen from Section 2.8 that the quantities $Q_j(\Psi; \Psi^{(k)})$ are already available from the E-step of the EM algorithm. Hence it is convenient to approximate the observed information matrix in terms of these quantities, or effectively, the $s_j(\Psi^{(k)})$. The latter can be viewed as the incomplete-data score statistic for the j th interval \mathcal{W}_j , and so corresponds to $s(\mathbf{w}_j; \Psi^{(k)})$, the score statistic for the single observation \mathbf{w}_j in the case of ungrouped data.

Corresponding to the expression (4.40) for the so-called empirical information matrix for i.i.d. observations available in ungrouped form, we can approximate the observed information matrix for grouped i.i.d. data by

$$\mathbf{I}_{e,g}(\Psi^{(k)}; \mathbf{y}) = \sum_{j=1}^r n_j s_j(\Psi^{(k)}) \mathbf{s}_j^T(\Psi^{(k)}) - n \bar{s}(\Psi^{(k)}) \bar{s}^T(\Psi^{(k)}), \quad (4.45)$$

where

$$\bar{s}(\Psi^{(k)}) = \sum_{j=1}^r n_j s_j(\Psi^{(k)})/n. \quad (4.46)$$

We can write (4.45) as

$$\mathbf{I}_{e,g}(\Psi^{(k)}; \mathbf{y}) = \sum_{j=1}^r n_j \{s_j(\Psi^{(k)}) - \bar{s}(\Psi^{(k)})\} \{s_j(\Psi^{(k)}) - \bar{s}(\Psi^{(k)})\}^T, \quad (4.47)$$

demonstrating that it is always positive definite. Note that in the presence of truncation, $\bar{s}(\Psi)$ is not zero at the MLE $\hat{\Psi}$. From (2.78), it can be seen that in the case of truncation the score statistic at any point Ψ_o can be expressed as

$$\mathbf{S}(\mathbf{y}; \Psi) = \sum_{j=1}^r \{n_j - \frac{n P_j(\Psi)}{P(\Psi)}\} s_j(\Psi). \quad (4.48)$$

As $\mathbf{S}(\mathbf{y}; \hat{\Psi}) = \mathbf{0}$, it implies that

$$\bar{s}(\hat{\Psi}) = \sum_{j=1}^r \{P_j(\hat{\Psi})/P(\hat{\Psi})\} s_j(\hat{\Psi}). \quad (4.49)$$

The use of (4.47) as an approximation to the observed matrix is developed in Jones and McLachlan (1992). It can be confirmed that $\mathbf{I}_{e,g}(\hat{\Psi}; \mathbf{y})/n$ is a consistent estimator of the expected information matrix $\mathbf{i}(\Psi)$ contained in a single observation. The inverse of $\mathbf{I}_{e,g}(\hat{\Psi}; \mathbf{y})$ provides an approximation to the covariance matrix of the MLE $\hat{\Psi}$, while

$\mathbf{I}_{e,g}(\Psi^{(k)}; \mathbf{y})$ provide an approximation to $\mathbf{I}(\Psi^{(k)}; \mathbf{y})$ for use, say, in the Newton-Raphson method of computing the MLE.

The use of $\mathbf{I}_{e,g}(\hat{\Psi}; \mathbf{y})$ can be justified in the same way that the use of the empirical observation matrix was in Section 4.3 for ungrouped i.i.d. data. For example, corresponding to the relationship (4.47) between the observed information matrix and its empirical analog, we have that

$$\mathbf{I}(\hat{\Psi}; \mathbf{y}) = \mathbf{I}_{e,g}(\hat{\Psi}; \mathbf{y}) + \mathbf{R}(\hat{\Psi}), \quad (4.50)$$

where

$$\begin{aligned} \mathbf{R}(\Psi) &= -\sum_{j=1}^r \{n_j/P_j(\Psi)\} \partial^2 P_j(\Psi)/\partial\Psi\partial\Psi^T \\ &\quad + \{n/P(\Psi)\} \partial^2 P(\Psi)/\partial\Psi\partial\Psi^T \end{aligned}$$

has zero expectation.

To verify (4.50), we have from differentiation of the (incomplete-data) log likelihood that

$$\begin{aligned} \mathbf{I}(\Psi; \mathbf{y}) &= \sum_{j=1}^r n_j \{\partial \log P_j(\Psi)/\partial\Psi\} \{\partial \log P_j(\Psi)/\partial\Psi\}^T \\ &\quad - n \{\partial \log P(\Psi)/\partial\Psi\} \{\partial \log P(\Psi)/\partial\Psi\}^T \\ &\quad + \mathbf{R}(\Psi). \end{aligned} \quad (4.51)$$

Now from (2.69) and (4.44),

$$\partial \log P_j(\Psi)/\partial\Psi = s_j(\Psi) \quad (4.52)$$

and

$$\partial \log P(\Psi)/\partial\Psi = \sum_{j=1}^r \{P_j(\Psi)/P(\Psi)\} s_j(\Psi).$$

Since $\mathbf{S}(\mathbf{y}; \hat{\Psi}) = \mathbf{0}$, we have from (2.78) that

$$\begin{aligned} \partial \log P(\hat{\Psi})/\partial\Psi &= \sum_{j=1}^r \{P_j(\hat{\Psi})/P(\hat{\Psi})\} s_j(\hat{\Psi}) \\ &= \sum_{j=1}^r (n_j/n) s_j(\hat{\Psi}) \\ &= \bar{s}(\hat{\Psi}). \end{aligned} \quad (4.53)$$

On using (4.52) and (4.53) in (4.51), we obtain the result (4.50).

4.4.2 Example 4.3: Grouped Data from an Exponential Distribution

We suppose that the random variable W has an exponential distribution with mean μ , as specified by (1.45). The sample space \mathcal{W} of W is partitioned into v mutually exclusive intervals \mathcal{W}_j of equal length d , where

$$\mathcal{W}_j = [w_{j-1}, w_j) \quad (j = 1, \dots, v)$$

and $w_0 = 0$ and $w_v = \infty$.

The problem is to estimate μ on the basis of the observed frequencies

$$\mathbf{y} = (n_1, \dots, n_v)^T,$$

where n_j denotes the number of observations on W falling in the j th interval \mathcal{W}_j ($j = 1, \dots, v$), and where

$$n = \sum_{j=1}^v n_j.$$

In Section 2.8, we considered ML estimation from grouped and truncated data in the general case. The present example is simpler in that there is no truncation. Proceeding as in Section 2.8, we declare the complete-data vector \mathbf{x} to be

$$\mathbf{x}_j = (\mathbf{y}^T, \mathbf{w}_1^T, \dots, \mathbf{w}_v^T)^T,$$

where

$$\mathbf{w}_j = (w_{j1}, \dots, w_{jn_j})^T$$

contains the n_j unobservable individual observations on W that fell in the j th interval \mathcal{W}_j ($j = 1, \dots, v$).

The complete-data log likelihood is given by

$$\begin{aligned} \log L_c(\mu) &= \sum_{j=1}^v \sum_{l=1}^{n_j} \log f(w_{jl}; \mu), \\ &= -n \log \mu - \sum_{j=1}^v \sum_{l=1}^{n_j} w_{jl}/\mu \end{aligned} \quad (4.54)$$

on substituting for $f(w_{jk}; \mu)$ from (1.45). Then

$$Q(\mu; \mu^{(k)}) = \sum_{j=1}^v n_j^{(k)} Q_j(\mu; \mu^{(k)}),$$

where

$$Q_j(\mu; \mu^{(k)}) = E_{\mu^{(k)}} \{ \log f(W; \mu) \mid W \in \mathcal{W}_j \}.$$

In the present case of an exponential distribution,

$$\begin{aligned} Q_j(\mu; \mu^{(k)}) &= E_{\mu^{(k)}} \{ (-\log \mu - \mu^{-1} W) \mid W \in \mathcal{W}_j \} \\ &= -\log \mu - \mu^{-1} w_j^{(k)}, \end{aligned} \quad (4.55)$$

where

$$w_j^{(k)} = E_{\mu^{(k)}} (W \mid W \in \mathcal{W}_j) \quad (4.56)$$

for $j = 1, \dots, v$.

The conditional expectation in (4.56) can be calculated to give

$$w_j^{(k)} = (c_j^{(k)} \mu^{(k)})^{-1} \int_{w_{j-1}}^{w_j} w \exp(-w/\mu^{(k)}) dw, \quad (4.57)$$

where

$$\begin{aligned} c_j^{(k)} &= \int_{w_{j-1}}^{w_j} \mu^{(k)-1} \exp(-w/\mu^{(k)}) dw \\ &= \exp(-w_{j-1}/\mu^{(k)}) \{1 - \exp(-d/\mu^{(k)})\}. \end{aligned} \quad (4.58)$$

On performing the integration in (4.57), it is easily seen that

$$w_j^{(k)} = \mu^{(k)} + a_j^{(k)}, \quad (4.59)$$

where

$$a_j^{(k)} = w_{j-1} - d/\{\exp(d/\mu^{(k)}) - 1\}. \quad (4.60)$$

On maximizing $Q(\mu; \mu^{(k)})$ with respect to μ , it follows that $\mu^{(k+1)}$ is given by

$$\begin{aligned} \mu^{(k+1)} &= \sum_{j=1}^v (n_j/n) w_j^{(k)} \\ &= \mu^{(k)} + \bar{a}^{(k)}, \end{aligned} \quad (4.61)$$

where

$$\bar{a}^{(k)} = \sum_{j=1}^v (n_j/n) a_j^{(k)}.$$

This completes the implementation of the $(k+1)$ th iteration of the EM algorithm. We now consider the application of a modified Newton-Raphson method to this problem. From (2.71), the (incomplete-data) score statistic at $\mu = \mu^{(k)}$ can be computed as

$$\begin{aligned} S(\mathbf{y}; \mu^{(k)}) &= [\partial Q(\mu; \mu^{(k)})/\partial \mu]_{\mu=\mu^{(k)}} \\ &= \sum_{j=1}^v n_j [\partial Q_j(\mu; \mu^{(k)})/\partial \mu]_{\mu=\mu^{(k)}} \\ &= \sum_{j=1}^v n_j s_j(\mu^{(k)}), \end{aligned} \quad (4.62)$$

where

$$s_j(\mu^{(k)}) = [\partial Q_j(\mu; \mu^{(k)})/\partial \mu]_{\mu=\mu^{(k)}}.$$

On differentiating (4.55) and noting (4.59), we have that

$$\begin{aligned} s_j(\mu^{(k)}) &= [-\mu^{-1} + \mu^{-2} w_j^{(k)}]_{\mu=\mu^{(k)}} \\ &= a_j^{(k)} / \mu^{(k)2} \quad (j = 1, \dots, v). \end{aligned} \quad (4.63)$$

On using (4.63) in (4.62), we have

$$\begin{aligned} S(\mathbf{y}; \mu^{(k)}) &= \sum_{j=1}^v n_j a_j^{(k)} / \mu^{(k)2} \\ &= n \bar{a}^{(k)} / \mu^{(k)2}. \end{aligned} \quad (4.64)$$

Corresponding to the use of the so-called empirical covariance matrix to approximate the observed matrix in the ungrouped case of i.i.d. data, Meilijson (1989) considers the Newton-Raphson method with the information $I(\mu^{(k)}; \mathbf{y})$ after the k th iteration approximated by

$$I_{e,g}(\mu^{(k)}; \mathbf{y}) = \sum_{j=1}^v n_j \{s_j(\mu^{(k)})\}^2 - n^{-1} \{S(\mathbf{y}; \mu^{(k)})\}^2. \quad (4.65)$$

The Newton-Raphson method with the modification (4.65) leads to $\mu^{(k+1)}$ being given by

$$\begin{aligned} \mu^{(k+1)} &= \mu^{(k)} + S(\mathbf{y}; \mu^{(k)})/I_{e,g}(\mu^{(k)}; \mathbf{y}) \\ &= \mu^{(k)} + h^{(k)}, \end{aligned} \quad (4.66)$$

where

$$h^{(k)} = \frac{\bar{a}^{(k)} \mu^{(k)} \mu^{(k)}}{\sum_{j=1}^v (n_j/n) a_j^{(k)2} - \bar{a}^{(k)2}}. \quad (4.67)$$

Meilijson (1989) contrasts the convergence of $\mu^{(k)}$ as given by (4.61) under the EM algorithm with $\mu^{(k)}$ as given by (4.66) with the modified Newton-Raphson method on various data sets. He demonstrates the superior global convergence when started from any positive value of μ , while the modified Newton-Raphson method diverges when started too close to zero and oscillates between two values on significantly bimodal data. The fit of the model is found to influence the convergence of the modified Newton-Raphson method, and the intervals that of the EM algorithm. When the model fit is good, the former method should be fast converging. When the intervals are sparse, the observed frequencies barely predict the complete-data statistic, and so the convergence of the EM algorithm is slow.

Since

$$S(\mathbf{y}; \hat{\mu}) = 0,$$

we have from (4.63) and (4.65) that

$$I_{e,g}(\hat{\mu}; \mathbf{y}) = n\hat{\mu}^{-4} \sum_{j=1}^v (n_j/n) \{a_j(\hat{\mu})\}^2, \quad (4.68)$$

where

$$a_j(\hat{\mu}) = w_{j-1} - d/\{\exp(d/\hat{\mu}) - 1\}.$$

The variance of the MLE $\hat{\mu}$ can be approximated by

$$\text{var}_{\mu}(\hat{\mu}) \approx 1/I_{e,g}(\hat{\mu}; \mathbf{y}).$$

4.5 SUPPLEMENTED EM ALGORITHM

4.5.1 Definition

The methods presented in the previous section are of course applicable only in the specialized case of data arising from i.i.d. observations. For the general case, Meng and Rubin (1989, 1991) define a procedure that obtains a numerically stable estimate of the asymptotic covariance matrix of the EM-computed estimate, using only the code for computing the

complete-data covariance matrix, the code for the EM algorithm itself, and the code for standard matrix operations. In particular, neither likelihoods, nor partial derivatives of log likelihoods need to be evaluated.

The basic idea is to use the fact that the rate of convergence is governed by the fraction of the missing information to find the increased variability due to missing information to add to the assessed complete-data covariance matrix. Meng and Rubin (1991) refer to the EM algorithm with their modification for the provision of the asymptotic covariance matrix as the Supplemented EM algorithm.

In his discussion of the DLR paper, Smith (1977) notes the possibility of obtaining the asymptotic variance in single parameter cases by using the rate of convergence r of the EM algorithm. He gives the expression

$$v = v_c / (1 - r), \quad (4.69)$$

where v and v_c denote the asymptotic variance of the maximum likelihood estimator based on the observed (incomplete) and complete data, respectively.

The expression (4.69) can be written in the form

$$v = v_c + \Delta v, \quad (4.70)$$

where

$$\Delta v = \{r / (1 - r)\} v_c$$

is the increase in the variance due to not observing the data in \mathbf{z} . Meng and Rubin (1991) extend this result to the multivariate case of $d > 1$ parameters.

Let \mathbf{V} denote the asymptotic covariance matrix of the MLE $\hat{\Psi}$. Then analogous to (4.70), Meng and Rubin (1991) show that

$$\mathbf{I}^{-1}(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y}) + \Delta \mathbf{V}, \quad (4.71)$$

where

$$\Delta \mathbf{V} = \{\mathbf{I}_d - \mathbf{J}(\hat{\Psi})\}^{-1} \mathbf{J}(\hat{\Psi}) \mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y})$$

and $\mathbf{J}(\Psi)$ is defined by (3.68). Hence the diagonal elements of $\Delta \mathbf{V}$ give the increases in the asymptotic variances of the components of $\hat{\Psi}$ due to missing data.

To derive the result (4.71), note that from (3.50),

$$\begin{aligned} \mathbf{I}(\Psi; \mathbf{y}) &= \mathcal{I}_c(\Psi; \mathbf{y}) - \mathcal{I}_m(\Psi; \mathbf{y}) \\ &= \mathcal{I}_c(\Psi; \mathbf{y}) \{ \mathbf{I}_d - \mathcal{I}_c^{-1}(\Psi; \mathbf{y}) \mathcal{I}_m(\Psi; \mathbf{y}) \}. \end{aligned}$$

From (3.72),

$$\mathbf{J}(\hat{\Psi}) = \mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y}) \mathcal{I}_m(\hat{\Psi}; \mathbf{y})$$

for an EM sequence satisfying (3.71). Hence the observed information matrix $\mathbf{I}(\hat{\Psi}; \mathbf{y})$ can be expressed as

$$\mathbf{I}(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c(\hat{\Psi}; \mathbf{y}) \{ \mathbf{I}_d - \mathbf{J}(\hat{\Psi}) \}, \quad (4.72)$$

which on inversion, yields

$$\begin{aligned} \mathbf{I}^{-1}(\hat{\Psi}; \mathbf{y}) &= \{ \mathbf{I}_d - \mathbf{J}(\hat{\Psi}) \}^{-1} \mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y}) \\ &= [\mathbf{I}_d + \{ \mathbf{I}_d - \mathbf{J}(\hat{\Psi}) \}^{-1} \mathbf{J}(\hat{\Psi})] \mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y}) \\ &= \mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y}) + \{ \mathbf{I}_d - \mathbf{J}(\hat{\Psi}) \}^{-1} \mathbf{J}(\hat{\Psi}) \mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y}), \end{aligned} \quad (4.73)$$

thus establishing (4.71).

Jamshidian and Jennrich (2000) give a derivation of the result (4.72), using calculations similar to those in the derivation by Oakes (1999) of an expression for the observed information matrix. These calculations are to be described in Section 4.7.3.

It can be seen that in order to use (4.72) to compute the observed information matrix, the conditional expected complete-data information matrix $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$ and the Jacobian matrix $\mathbf{J}(\hat{\Psi})$ need to be calculated. For a wide class of problems where the complete-data density is from the regular exponential family, the evaluation of $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$ is readily facilitated by standard complete-data computations. As discussed in Section 4.2.4, if the complete-data density $g_c(\mathbf{x}; \Psi)$ belongs to the regular exponential family, then

$$\mathcal{I}_c(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c(\hat{\Psi}).$$

Concerning the calculation of $\mathbf{J}(\hat{\Psi})$, Meng and Rubin (1991) demonstrate how $\mathbf{J}(\hat{\Psi})$ can be readily obtained by using only EM code. The procedure amounts to numerically differentiating the EM map $\mathbf{M}(\Psi)$, as we now describe.

4.5.2 Calculation of $\mathbf{J}(\hat{\Psi})$ via Numerical Differentiation

We consider here the calculation of $\mathbf{J}(\hat{\Psi})$. The component-wise convergence rates of the EM sequence $\{\Psi^{(k)}\}$,

$$\lim_{k \rightarrow \infty} (\Psi_i^{(k+1)} - \Psi_i^{(k)}) / (\Psi_i^{(k)} - \Psi_i^{(k-1)}) \quad (i = 1, \dots, d),$$

provide only a few eigenvalues (in most cases, simply the largest eigenvalue) of $\mathbf{J}(\hat{\Psi})$, and not the matrix itself. However, as explained by Meng and Rubin (1991), each element of $\mathbf{J}(\hat{\Psi})$ is the component-wise rate of convergence of a “forced EM”. Let r_{ij} be the (i, j) th element of $\mathbf{J}(\hat{\Psi})$, and define $\Psi_{(j)}^{(k)}$ to be

$$\Psi_{(j)}^{(k)} = (\hat{\Psi}_1, \dots, \hat{\Psi}_{j-1}, \Psi_j^{(k)}, \hat{\Psi}_{j+1}, \dots, \hat{\Psi}_d)^T, \quad (4.74)$$

where $\Psi_j^{(k)}$ is the value of Ψ_j on the k th iteration in a subsequent application of the EM algorithm, as explained below.

By the definition of r_{ij} ,

$$\begin{aligned} r_{ij} &= \partial M_i(\hat{\Psi}) / \partial \Psi_j \\ &= \lim_{\Psi_j \rightarrow \hat{\Psi}_j} \frac{M_i(\hat{\Psi}_1, \dots, \hat{\Psi}_{j-1}, \Psi_j, \hat{\Psi}_{j+1}, \dots, \hat{\Psi}_d) - M_i(\hat{\Psi})}{\Psi_j - \hat{\Psi}_j} \\ &= \lim_{k \rightarrow \infty} \frac{M_i(\Psi_{(j)}^{(k)}) - \hat{\Psi}_i}{\Psi_j^{(k)} - \hat{\Psi}_j} \\ &= \lim_{k \rightarrow \infty} r_{ij}^{(k)}, \end{aligned} \quad (4.75)$$

where

$$r_{ij}^{(k)} = \frac{M_i(\Psi_{(j)}^{(k)}) - \hat{\Psi}_i}{\Psi_j^{(k)} - \hat{\Psi}_j}. \quad (4.76)$$

Because $M(\Psi)$ is implicitly defined by the output of the E- and M-steps, all quantities in (4.76) can be obtained using only the code for the EM algorithm. Meng and Rubin (1991) suggested the following algorithm for computing the $r_{ij}^{(k)}$ after $\hat{\Psi}$ has been found. Firstly, run a sequence of iterations of the EM algorithm, starting from a point that is not equal to $\hat{\Psi}$ in any component. After the k th iteration, compute $\Psi_{(j)}^{(k)}$ from (4.74) and, treating it as the current estimate of Ψ , run one iteration of the EM algorithm to obtain $M_i(\Psi_{(j)}^{(k)})$ and, subsequently from (4.76), the ratio $r_{ij}^{(k)}$ for each i ($i = 1, \dots, d$). This is done for each j ($j = 1, \dots, d$).

The r_{ij} are obtained when the sequence $r_{ij}^{(k^*)}, r_{ij}^{(k^*+1)}, \dots$ is stable for some k^* . This process may result in using different values of k^* for different r_{ij} elements.

As for numerical accuracy, it is almost always safe to use the original EM starting values as the initial values for the Supplemented EM algorithm in computing $J(\hat{\Psi})$. This choice does not require any additional work, but may result in some unnecessary iterations because the original starting values may be far from the MLE. Meng and Rubin (1991) suggest using a suitable iterate of the original EM sequence (for example, the second) or two complete-data standard deviations from the MLE. The stopping criterion for the Supplemented EM sequence should be less stringent than that for the original EM sequence because the method for computing $J(\hat{\Psi})$ is essentially numerical differentiation of a function.

4.5.3 Stability

The observed information matrix $I(\hat{\Psi}; \mathbf{y})$ can be approximated by using methods such as those described by Carlin (1987) and Meilijson (1989), where the second-order partial derivatives of the log likelihood function or at least its gradient are calculated by numerical differentiation. However, as pointed out by Meng and Rubin (1991), these methods besides requiring the evaluation of this likelihood, are subject to the inaccuracies and difficulties of any numerical differentiation procedure with large matrices. On the other hand, the approach via the Supplemented EM algorithm, is typically more stable than pure numerical differentiation procedures, because the matrix differentiation is being added to an analytically obtained matrix, which is usually the dominant term. Thus the Supplemented EM algorithm is typically more stable than pure numerical differentiation procedures that are used to compute the whole covariance matrix.

More specifically, with the Supplemented EM algorithm, there is an automatic self-adjustment that Meng and Rubin (1991) explain as follows. When the fraction of missing data is high, the convergence of the EM algorithm is slow. As a consequence, it provides an excellent sequence of iterates from which the linear rate of convergence of the EM algorithm can be assessed, and thus the results produced by the Supplemented EM algorithm are typically quite accurate.

On the other hand, when the fraction of missing data is low and consequently the increases in the variances of the fitted parameters are relatively small, the term $\mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y})$ in (4.71), which is usually very accurately calculated, dominates the increases in the variances due to the missing data. Thus the Supplemented EM algorithm still provides a satisfactory assessment of the observed information matrix $I(\hat{\Psi}; \mathbf{y})$, even though it is not as accurate in assessing the increases in the variances as when convergence of the EM algorithm is slow.

4.5.4 Monitoring Convergence

The matrix

$$\mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y}) + \Delta \mathbf{V}$$

is not numerically constrained to be symmetric, even though it is mathematically symmetric. The asymmetry can arise because of inaccuracies in computing either $\mathbf{J}(\hat{\Psi})$ or $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$. For exponential families, $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$ is typically very accurately computed, whereas for nonexponential families, its accuracy depends on large-sample approximations based on linearization methods. In contrast, the accuracy of $\mathbf{J}(\hat{\Psi})$ is determined by the accuracy of the EM algorithm itself, which typically is excellent when both the E- and M-steps are simple calculations and is adequate in most cases.

If the matrix $\mathbf{I}_d - \mathbf{J}(\hat{\Psi})$ is (numerically) symmetric, but not positive semidefinite, then it indicates that the EM algorithm has not converged to a (local) maximum, but rather to a saddle point. In the latter case the EM algorithm should be rerun, starting near the last iterate but perturbed in the direction of the eigenvector corresponding to the most negative eigenvalue of $\mathbf{I}(\hat{\Psi}; \mathbf{y})$. In this sense, the Supplemented EM algorithm can also be used to monitor the convergence of the EM algorithm to a local maximum, which cannot be detected by monitoring the increase in the likelihood.

The matrix $\mathbf{I}_d - \mathbf{J}(\hat{\Psi})$ can be nearly singular when the convergence of the EM algorithm is extremely slow; that is, when the largest eigenvalue of $\mathbf{J}(\hat{\Psi})$ is very close to one. Statistically this implies that $L(\Psi)$ is very flat along some directions and thus that $\mathbf{I}(\hat{\Psi}; \mathbf{y})$ is nearly singular. This is a feature of the likelihood function and the data and not a problem created by the Supplemented EM algorithm. Even in these situations, the Supplemented EM algorithm can be very helpful in identifying directions with little information by proceeding as follows.

First obtain the matrix

$$\mathbf{P} = \mathcal{I}_c(\hat{\Psi}; \mathbf{y})\{\mathbf{I}_d - \mathbf{J}(\hat{\Psi})\}. \quad (4.77)$$

As discussed above, any lack of (numerical) symmetry in \mathbf{P} indicates the existence of programming errors or lack of convergence. Assuming (numerical) symmetry of \mathbf{P} , standard matrix operations can be used to find the spectral decomposition of \mathbf{P} ,

$$\mathbf{P} = \mathbf{F}^T \text{diag}(\lambda_1, \dots, \lambda_d) \mathbf{F}, \quad (4.78)$$

where $\lambda_1 \geq \lambda_2 \dots \geq \lambda_d$ are the eigenvalues of \mathbf{P} and where the rows of \mathbf{F} are the orthonormalized eigenvectors corresponding to $\lambda_1, \dots, \lambda_d$. If λ_d is identified as being (numerically) negative, it indicates that the EM algorithm has not converged to a local maximum, but to a saddle point, and that it should be continued in the direction corresponding to λ_d . Otherwise, the eigenvalues can be used to indicate those components of $\mathbf{F}\Psi$ about which the observed data contains little or no information.

4.5.5 Difficulties of the SEM Algorithm

There are, however, some difficulties in the use of the SEM algorithm. They are:

1. It seems to be susceptible to numerical inaccuracies and instability, especially in high-dimensional settings (Baker, 1992; McCulloch, 1998; Segal et al., 1994).
2. It requires code for complete-data information matrix, which is not always available, if the model is complicated (Baker, 1992).

3. It can be much more expensive (Belin and Rubin, 1995).

Jamshidian and Jennrich (2000) discuss various issues concerning approximately obtaining observed information matrix by numerical differentiation. They develop mainly forward difference (FDM) and Richardson extrapolation (REM) methods as alternatives to the SEM algorithm and discuss several examples.

4.5.6 Example 4.4: Univariate Contaminated Normal Data

This example is considered by Meng and Rubin (1991) and prior to that by Little and Rubin (2002, Section 12.2). They assume that the observed data are given by

$$\mathbf{y} = (w_1, \dots, w_n)^T,$$

where w_1, \dots, w_n denote the realizations of a random sample of size n from the univariate contaminated normal model,

$$f(w; \Psi) = \pi_1\phi(w; \mu, \sigma^2) + \pi_2\phi(w; \mu, c\sigma^2), \quad (4.79)$$

and $\pi_1(0 < \pi_1 < 1)$ and $c(c > 0)$ are both known; $\pi_2 = 1 - \pi_1$. The model (4.79) represents a mixture in known proportions of two univariate normal densities with a common mean and with variances in a known ratio. It is therefore a special case of the general univariate normal mixture model considered in Section 2.7.1. On specializing the equations there for the model (4.79), we have that on the $(k+1)$ th iteration of the EM algorithm, $\Psi^{(k+1)} = (\mu^{(k+1)}, \sigma^{(k+1)^2})^T$ satisfies

$$\mu^{(k+1)} = \frac{\sum_{j=1}^n \{z_{ij}^{(k)} + c^{-1}z_{2j}^{(k)}\}w_j}{\sum_{j=1}^n \{z_{1j}^{(k)} + c^{-1}z_{2j}^{(k)}\}}$$

and

$$\begin{aligned} \sigma^{(k+1)^2} = & \sum_{j=1}^n \{z_{ij}^{(k)}(w_j - \mu^{(k+1)})^2 \\ & + z_{2j}^{(k)}c^{-1}(w_j - \mu^{(k+1)})^2\}/n, \end{aligned}$$

where

$$z_{ij}^{(k)} = \tau_i(w_j; \Psi^{(k)}),$$

and where

$$\tau_1(w_j; \Psi) = \frac{\pi_1\phi(w_j; \mu, \sigma^2)}{\pi_1\phi(w_j; \mu, \sigma^2) + \pi_2\phi(w_j; \mu, c\sigma^2)}$$

and

$$\tau_2(w_j; \Psi) = 1 - \tau_1(w_j; \Psi) \quad (j = 1, \dots, n).$$

To illustrate the Supplemented EM algorithm, Meng and Rubin (1991) performed a simulation with

$$\mu = 0, \quad \sigma^2 = 1, \quad c = 0.5, \quad \pi_1 = 0.9, \quad \text{and } n = 100.$$

Meng and Rubin (1991) actually work with the parameter vector

$$(\mu, \log \sigma^2)^T \quad (4.80)$$

in order to improve the normal approximation to the log likelihood function. In the remainder of this section, we therefore let Ψ be the vector (4.80).

Table 4.1 gives the EM output with initial values

$$\mu^{(0)} = \bar{w}$$

and

$$\sigma^{(0)2} = s^2 / (\pi_1 + c\pi_2),$$

where \bar{w} and s^2 denote the sample mean and (bias-corrected) sample variance of the observed data w_1, \dots, w_n .

In this table,

$$d_1^{(k+1)} = \mu^{(k+1)} - \hat{\mu}$$

and

$$d_2^{(k+1)} = \log \sigma^{(k+1)2} - \log \hat{\sigma}^2.$$

The first four rows for Table 4.2 give the corresponding output for $r_{ij}^{(k)}$ ($i, j = 1, 2$) for $k = 0, 1, 2$, and 3, obtained by the Supplemented EM algorithm starting from $\Psi^{(0)}$ and taking $\hat{\Psi} = \Psi^{(6)}$. The last row gives the true values of r_{ij} ($i, j = 1, 2$) obtained by direct computation using analytical expressions.

As noted by Meng and Rubin (1991), it can be seen from Table 4.2 that using only six iterations initially to form $\hat{\Psi}$, we can approximate the r_{ij} very well by $r_{ij}^{(k)}$ for “moderate” k .

Table 4.1 Results of EM Algorithm for Contaminated Normal Example.

k	$\mu^{(k)}$	$d_1^{(k)}$	$d_1^{(k+1)}/d_1^{(k)}$	$\log \sigma^{(k)2}$	$d_2^{(k)}$	$d_2^{(k+1)}/d_2^{(k)}$
0	0.1986692	-0.00061984	0.01587	0.22943020	0.00923662	0.03502
1	0.1992569	-0.00003215	0.04890	0.22051706	0.00032348	0.03496
2	0.1992874	-0.00000157	0.04708	0.22020489	0.00001131	0.03497
3	0.1992889	-0.00000007	0.04590	0.22019397	0.00000040	0.03498
4	0.1992890	-0.00000000	0.04509	0.22019359	0.00000001	0.03499
5	0.1992890	-0.00000000	0.04443	0.22019358	0.00000000	0.03497
6	0.1992890	-0.00000000	0.04223	0.22019358	0.00000000	0.03386

Source: Adapted from Meng and Rubin (1991), with permission of the Journal of the American Statistical Association

From Table 4.2, we can take

$$\mathbf{J}(\hat{\Psi}) = \begin{pmatrix} 0.04252 & -0.00064 \\ -0.00112 & 0.03492 \end{pmatrix}. \quad (4.81)$$

Table 4.2 Results of SEM Iterations for \mathbf{J} Matrix in Contaminated Normal Example.

Iteration (k)	$r_{11}^{(k)}$	$r_{12}^{(k)}$	$r_{21}^{(k)}$	$r_{22}^{(k)}$
0	0.04251717	-0.00063697	-0.0012649	0.03494620
1	0.04251677	-0.00063697	-0.0012375	0.03492154
2	0.04251674	-0.00063666	-0.0012359	0.03492066
3	0.04251658	-0.00063663	-0.0012333	0.03492059
True	0.04251675	-0.00063666	-0.0012360	0.03492063

Source: Adapted from Meng and Rubin (1991), with permission of the Journal of the American Statistical Association

Concerning the computation of $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$, we have that the first and second diagonal elements of $\mathbf{I}_c(\Psi; \mathbf{x})$ are given by

$$-\partial^2 \log L_c(\Psi) / \partial \mu^2 = \sigma^{-2} \sum_{j=1}^n (z_{1j} + c^{-1} z_{2j}) \quad (4.82)$$

and

$$-\partial^2 \log L_c(\Psi) / \partial \Psi_2^2 = \frac{1}{2} n, \quad (4.83)$$

respectively, where $\Psi_2 = \log \sigma^2$. The z_{ij} are the zero-one component indicator variables defined to be one or zero according as the j th observation arises or does not arise from the i th component of the mixture model (4.79). The common value of the two off-diagonal elements of $\mathbf{I}_c(\Psi; \mathbf{x})$ is not given here as it obviously will have zero conditional expectation.

On taking the conditional expectation of (4.82) given the observed data \mathbf{y} , we have that

$$\begin{aligned} E_{\Psi} \{-\partial^2 \log L / \partial \mu^2 \mid \mathbf{y}\} &= \sigma^{-2} \sum_{j=1}^n \{\tau_1(w_j; \Psi) + c^{-1} \tau_2(w_j; \Psi)\} \\ &= \sigma^{-2} \xi(\Psi), \end{aligned}$$

say. It follows that

$$\mathcal{I}_c(\hat{\Psi}; \mathbf{y}) = \text{diag}(\hat{\sigma}^{-2} \xi(\hat{\Psi}), \frac{1}{2} n) \quad (4.84)$$

We have seen in Section 4.2 that if the complete-data density belongs to the regular exponential family, then

$$\mathcal{I}_c(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c(\hat{\Psi}). \quad (4.85)$$

However, for this example, the complete-data density is from an irregular exponential family, and so (4.85) does not necessarily hold. Indeed, it can be confirmed that (4.85) does not hold, since

$$\mathcal{I}_c(\Psi) = \text{diag}\{n(\pi_1 + c^{-1} \pi_2), \frac{1}{2} n\}.$$

Of course, if π_1 were unknown, the complete-data density belongs to the regular exponential family, and it can be easily confirmed then that (4.85) holds.

From (4.84), we have that

$$\begin{aligned} \mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y}) &= n^{-1} \begin{pmatrix} \hat{\sigma}^2 / \xi(\hat{\Psi}) & 0 \\ 0 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 0.01133 & 0 \\ 0 & 2 \end{pmatrix}, \end{aligned} \quad (4.86)$$

since $n\xi(\hat{\Psi}) = 109.98$. Thus using (4.81) and (4.86) in (4.73), we have

$$\mathbf{I}^{-1}(\hat{\Psi}; \mathbf{y}) = \begin{pmatrix} 0.01184 & -0.00001 \\ -0.00001 & 0.02072 \end{pmatrix}.$$

The symmetry of the resulting matrix indicates numerical accuracy, as discussed earlier in this section.

It can be seen from the fourth and seventh columns of Table 4.1 that the iterates $\Psi_1 = \mu$ and $\Psi_2 = \log \sigma^2$ converge at different rates, corresponding to two different eigenvalues of $\mathbf{J}(\hat{\Psi})$. This special feature occurs because the MLE's of Ψ_1 and Ψ_2 are asymptotically independent for both complete- and incomplete-data problems.

4.5.7 Example 4.5: Bivariate Normal Data with Missing Values

We now consider the situation where there is no missing information on some components of Ψ . Partition Ψ as

$$\Psi = (\Psi_1^T, \Psi_2^T)^T, \quad (4.87)$$

where there is no missing information on the d_1 -dimensional subvector Ψ_1 of Ψ .

Corresponding to the partition (4.87) of Ψ , we now partition $\mathbf{J}(\Psi)$ and $\mathcal{I}_c^{-1}(\Psi; \mathbf{y})$ as

$$\mathbf{J}(\Psi) = \begin{pmatrix} \mathbf{J}_{11}(\Psi) & \mathbf{J}_{12}(\Psi) \\ \mathbf{J}_{21}(\Psi) & \mathbf{J}_{22}(\Psi) \end{pmatrix}$$

and

$$\mathcal{I}_c^{-1}(\Psi; \mathbf{y}) = \begin{pmatrix} \mathbf{G}_{11}(\Psi; \mathbf{y}) & \mathbf{G}_{12}(\Psi; \mathbf{y}) \\ \mathbf{G}_{21}(\Psi; \mathbf{y}) & \mathbf{G}_{22}(\Psi; \mathbf{y}) \end{pmatrix}.$$

The EM sequence $\{\Psi_1^{(k)}\}$ for Ψ_1 will converge in one iteration for Ψ_1 regardless of the starting value, with the result that the corresponding components of $\mathbf{M}(\Psi)$ will be a constant with zero derivative. That is, $\mathbf{J}_{11}(\Psi)$ and $\mathbf{J}_{21}(\Psi)$ are null matrices.

Meng and Rubin (1991) show that the observed information matrix for $\hat{\Psi} = (\hat{\Psi}_1^T, \hat{\Psi}_2^T)^T$ is given by

$$\mathbf{I}(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y}) + \Delta \mathbf{V},$$

where

$$\Delta \mathbf{V} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Delta \mathbf{V}_{22} \end{pmatrix}$$

and where

$$\Delta \mathbf{V}_{22} = \{\mathbf{I}_d - \mathbf{J}_{22}(\hat{\Psi})\}^{-1} \mathbf{J}_{22}(\hat{\Psi}) \mathbf{A}(\hat{\Psi}; \mathbf{y}) \quad (4.88)$$

and

$$\mathbf{A}(\hat{\Psi}; \mathbf{y}) = \mathbf{G}_{22}(\hat{\Psi}; \mathbf{y}) - \mathbf{G}_{21}(\hat{\Psi}; \mathbf{y}) \mathbf{G}_{11}^{-1}(\hat{\Psi}; \mathbf{y}) \mathbf{G}_{12}(\hat{\Psi}; \mathbf{y}).$$

The submatrix $\mathbf{J}_{22}(\hat{\Psi})$ of $\mathbf{J}(\hat{\Psi})$ in (4.88) can be assessed by the same method used to assess $\mathbf{J}(\hat{\Psi})$, as explained in the last section.

To illustrate the application of the Supplemented EM algorithm with no missing information on some of the components of Ψ , Meng and Rubin (1991) consider ML estimation of

$$\Psi = (\mu_1, \log \sigma_1^2, \mu_2, \log \sigma_2^2, \zeta)^T$$

in the bivariate normal distribution from $n = 18$ bivariate observations as listed below. The value of the second variable for the last six cases is missing (indicated by ?).

$$\begin{array}{ll} \text{Variate 1:} & 8 \quad 6 \quad 11 \quad 22 \quad 14 \quad 17 \quad 18 \quad 24 \quad 19 \\ \text{Variate 2:} & 59 \quad 58 \quad 56 \quad 53 \quad 50 \quad 45 \quad 43 \quad 42 \quad 39 \end{array}$$

$$\begin{array}{ll} \text{Variate 1:} & 23 \quad 26 \quad 40 \quad 4 \quad 4 \quad 5 \quad 6 \quad 8 \quad 10 \\ \text{Variate 2:} & 38 \quad 30 \quad 27 \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \end{array}$$

We have arranged the components in Ψ so that

$$\Psi = (\Psi_1^T, \Psi_2^T)^T,$$

where there is no missing information on the subvector

$$\Psi_1 = (\mu_1, \log \sigma_1^2)^T.$$

Here

$$\zeta = \frac{1}{2} \log\{(1 + \rho)/(1 - \rho)\}$$

is Fisher's Z transformation of the correlation ρ .

As the first variate is fully observed, the MLE of Ψ_1 is simply the mean and log of the sample variance of the observations on it.

Using the Supplemented EM algorithm as described in Section 4.5.1, the submatrix $J_{22}(\hat{\Psi})$ of $J(\hat{\Psi})$ corresponding to the subvector

$$\Psi_2 = (\mu_2, \log \sigma_2^2, \zeta)^T$$

is obtained by Meng and Rubin (1991) to be

$$J_{22}(\hat{\Psi}) = \begin{pmatrix} 0.33333 & 1.44444 & -0.64222 \\ 0.05037 & 0.29894 & 0.01529 \\ -0.02814 & 0.01921 & 0.32479 \end{pmatrix}.$$

Since the complete-data density is from a regular exponential family, $\mathcal{I}_c(\hat{\Psi}; \mathbf{y})$ is equal to $\mathcal{I}_c(\hat{\Psi})$. From standard asymptotic results for the MLE of the vector of parameters of the bivariate normal distribution, we have that

$$\mathcal{I}_c(\Psi) = \begin{pmatrix} G_{11}(\Psi) & G_{12}(\Psi) \\ G_{21}(\Psi) & G_{22}(\Psi) \end{pmatrix}$$

where

$$G_{11}(\Psi) = \text{diag}\{n\sigma_1^{-2}(1 - \rho^2)^{-1}, (n/4)(2 - \rho^2)(1 - \rho^2)^{-1}\}^T,$$

$$G_{21}(\Psi) = \begin{pmatrix} -n\sigma_1^{-1}\sigma_2^{-1}\rho(1 - \rho^2)^{-1} & 0 & -(n/4)\rho^2(1 - \rho^2)^{-1} \\ 0 & 0 & -\frac{1}{2}n\rho \end{pmatrix},$$

and

$$G_{22}(\Psi) = \begin{pmatrix} n\sigma_2^{-2}(1 - \rho^2)^{-1} & 0 & 0 \\ 0 & (n/4)(2 - \rho^2)(1 - \rho^2)^{-1} & -\frac{1}{2}n\rho \\ 0 & -\frac{1}{2}n\rho & n(1 + \rho^2) \end{pmatrix},$$

and $G_{12}(\Psi) = G_{21}^T(\Psi)$.

On evaluation at $\Psi = \hat{\Psi}$, Meng and Rubin (1991) find that

$$\begin{aligned} G_{11}(\hat{\Psi}) &= \text{diag}(4.9741, 0.1111), \\ G_{12}(\hat{\Psi}) &= \begin{pmatrix} -5.0387 & 0 & 0 \\ 0 & 0.0890 & -0.0497 \end{pmatrix}, \\ G_{22}(\hat{\Psi}) &= \begin{pmatrix} 6.3719 & 0 & 0 \\ 0 & 0.1111 & -0.0497 \\ 0 & -0.0497 & 0.0556 \end{pmatrix}. \end{aligned}$$

Using the formula (4.88), Meng and Rubin (1991) find that

$$\Delta V_{22} = \begin{pmatrix} 1.0858 & 0.1671 & -0.0933 \\ 0.1671 & 0.0286 & -0.0098 \\ -0.0933 & -0.0098 & 0.0194 \end{pmatrix}.$$

The observed information matrix $I(\hat{\Psi}; \mathbf{y})$ is obtained by adding ΔV to $\mathcal{I}_c^{-1}(\hat{\Psi}; \mathbf{y})$. For example, the standard error of $\hat{\mu}_2$ is given by

$$(6.3719 + 1.0858)^{1/2} \cong 2.73.$$

4.6 BOOTSTRAP APPROACH TO STANDARD ERROR APPROXIMATION

The bootstrap was introduced by Efron (1979), who has investigated it further in a series of articles; see Efron and Tibshirani (1993) and the references therein. Efron (1994) considers the application of the bootstrap to missing-data problems. Over the past two or three decades, the bootstrap has become one of the most popular recent developments in statistics. Hence there now exists an extensive literature on it; see Chernick (2008), who has added a second bibliography to the first edition, containing about 1000 new references.

The bootstrap is a powerful technique that permits the variability in a random quantity to be assessed using just the data at hand. An estimate \hat{F} of the underlying distribution is formed from the observed sample. Conditional on the latter, the sampling distribution of the random quantity of interest with F replaced by \hat{F} , defines its so-called bootstrap distribution, which provides an approximation to its true distribution. It is assumed that \hat{F} has been so formed that the stochastic structure of the model has been preserved. Usually, it is impossible to express the bootstrap distribution in simple form, and it must be approximated by Monte Carlo methods whereby pseudo-random samples (bootstrap samples) are drawn from \hat{F} . There have been a number of papers written on improving the efficiency of the bootstrap computations with the latter approach. The nonparametric bootstrap uses $\hat{F} = \hat{F}_n$, the empirical distribution function of \mathbf{Y} formed from \mathbf{y} . If a parametric form is adopted for the distribution function of \mathbf{Y} , where Ψ denotes the vector of unknown parameters, then the parametric bootstrap uses an estimate $\hat{\Psi}$ formed from \mathbf{y} in place of Ψ . That is, if we write F as F_{Ψ} to signify its dependence on Ψ , then the bootstrap data are generated from $\hat{F} = F_{\hat{\Psi}}$.

Standard error estimation of $\hat{\Psi}$ may be implemented according to the bootstrap as follows:

Step 1. A new set of data, \mathbf{y}^* , called the bootstrap sample, is generated according to \hat{F} , an estimate of the distribution function of \mathbf{Y} formed from the original observed data \mathbf{y} .

That is, in the case where \mathbf{y} contains the observed values of a random sample of size n on a random vector \mathbf{W} , $\hat{\Psi}^*$ consists of the observed values of the random sample

$$\mathbf{W}_1^*, \dots, \mathbf{W}_n^* \stackrel{\text{iid}}{\sim} \hat{F}_{\mathbf{W}}, \quad (4.89)$$

where $\hat{F}_{\mathbf{W}}$ is held fixed at its observed value.

Step 2. The EM algorithm is applied to the bootstrap observed data \mathbf{y}^* to compute the MLE for this data set, $\hat{\Psi}^*$.

Step 3. The bootstrap covariance matrix of $\hat{\Psi}^*$ is given by

$$\text{cov}^*(\hat{\Psi}^*) = E^*[\{(\hat{\Psi}^* - E^*(\hat{\Psi}^*))\}\{(\hat{\Psi}^* - E^*(\hat{\Psi}^*))\}^T], \quad (4.90)$$

where E^* denotes expectation over the distribution of \mathbf{Y}^* specified by \hat{F} .

The bootstrap covariance matrix can be approximated by Monte Carlo methods. Steps (1) and (2) are repeated independently a number of times (say, B) to give B independent realizations of $\hat{\Psi}^*$, denoted by $\hat{\Psi}_1^*, \dots, \hat{\Psi}_B^*$. Then (4.90) can be approximated by the sample covariance matrix of these B bootstrap replications to give

$$\text{cov}^*(\hat{\Psi}^*) \approx \sum_{b=1}^B (\hat{\Psi}_b^* - \bar{\hat{\Psi}}^*)(\hat{\Psi}_b^* - \bar{\hat{\Psi}}^*)^T / (B - 1), \quad (4.91)$$

where

$$\bar{\hat{\Psi}}^* = \sum_{b=1}^B \hat{\Psi}_b^* / B. \quad (4.92)$$

The standard error of the i th element of $\hat{\Psi}$ can be estimated by the positive square root of the i th diagonal element of (4.91). It has been shown that 50 to 100 bootstrap replications are generally sufficient for standard error estimation (Efron and Tibshirani, 1993).

In Step 1 of the above algorithm, the nonparametric version of the bootstrap would take \hat{F} to be the empirical distribution function. Given that we are concerned here with ML estimation in the context of a parametric model, we would tend to use the parametric version of the bootstrap instead of the nonparametric version. Situations where we may still wish to use the latter include problems where the observed data are censored or are missing in the conventional sense. In these cases, the use of the nonparametric bootstrap avoids having to postulate a suitable model for the underlying mechanism that controls the censorship or the absence of the data.

4.7 BAKER'S, LOUIS', AND OAKES' METHODS FOR STANDARD ERROR COMPUTATION

4.7.1 Baker's Method for Standard Error Computation

Baker (1992) reviews methods for computing standard errors in the EM context and also develops a method for computing the observed information matrix in the case of categorical data. Jamshidian and Jennrich (2000) review more recent methods, including the Supplemented EM (SEM) algorithm of Meng and Rubin (1991), and suggest some newer methods based on numerical differentiation.

The expected information is generally more easily computed than the observed information matrix, especially in the case of categorical data, as in our example of Section 2.4. If the EM algorithm is used, then this matrix can be computed at the EM solution and inverted to get an estimate of the covariance matrix of the estimates. However, the expected information matrix is not as useful as the observed information matrix, the standard errors computed from the latter being more desirable for conditional, likelihood-based or Bayesian inference. On the other hand, the performance of resampling-based estimators of standard errors is not clear for even moderate size samples.

In simple models, the method of Hartley (1958) and Hartley and Hocking (1971) may be used for computing the observed information matrix. Here in each iteration the imputed missing values are substituted in the complete-data score vector to estimate the incomplete-data score vector. Using this and the EM iterates, a system of simultaneous equations are created from which the observed information matrix is computed.

Theoretically one may compute the asymptotic covariance matrix by inverting the observed or expected information matrix at the MLE. In practice, however, this may be tedious analytically or computationally, defeating one of the advantages of the EM approach. An alternative approach is to numerically differentiate the log likelihood function $\log L(\Psi)$ to obtain the Hessian. In a EM-aided differentiation approach, Meilija (1989) suggests perturbation of the incomplete-data score vector to compute the observed information matrix. Let $S(\mathbf{y}; \Psi)$ be the incomplete-data score vector. Perturb $\hat{\Psi}$ by adding a small amount $\epsilon > 0$ to the i th coordinate and compute $S(\mathbf{y}; \Psi)$ at the perturbed value $\tilde{\Psi}$ of $\hat{\Psi}$. Then the i th row of the Hessian is approximately

$$\{S(\mathbf{y}; \tilde{\Psi}) - S(\mathbf{y}; \hat{\Psi})\}/\epsilon.$$

This is carried out in turn for $i = 1, \dots, d$, the number of parameters.

4.7.2 Louis' Method of Standard Error Computation

The method of Louis (1982) requires only first and second derivatives of the complete-data log likelihood function, which are generally easier to work out than the corresponding derivatives of the incomplete-data log likelihood function. From (4.8), Louis' formula for the observed (incomplete-data) information matrix $I(\hat{\Psi}; \mathbf{y})$ is given by

$$I(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c(\hat{\Psi}; \mathbf{y}) - [\text{cov}_{\Psi} \{S_c(\mathbf{X}; \Psi) | \mathbf{y}\}]_{\Psi=\hat{\Psi}}. \quad (4.93)$$

Using the definition (1.41), the expected (conditional) complete-data information matrix $\mathcal{I}_c(\Psi; \mathbf{y})$ can be expressed as

$$\begin{aligned} \mathcal{I}_c(\Psi; \mathbf{y}) &= E_{\Psi} \{-\partial^2 \log L_c(\Psi)/\partial \Psi \partial \Psi^T | \mathbf{y}\} \\ &= -[\partial^2 Q(\Psi; \Psi_o)/\partial \Psi \partial \Psi^T]_{\Psi_o=\Psi}. \end{aligned} \quad (4.94)$$

This assumes we can interchange the order of differentiation and integration (expectation). This is assumed in the remainder of this section.

The result (4.93) can be used to compute the observed information matrix, which on inversion, gives an estimate of the covariance matrix of the MLE. The formula simplifies in the case of the multinomial model, but unfortunately only in cases where the second derivatives of the expected cell frequencies are all zero as is the case with our Example 1 of Section 2.4. We demonstrate Louis' method for this example in Section 4.7.4 below.

Wei and Tanner (1990a) develop a Monte Carlo version of Louis' method where the integration is replaced by a Monte Carlo procedure.

4.7.3 Oakes' Formula for Standard Error Computation

From (3.4), we have the identity in Ψ and $\Psi^{(k)}$,

$$\log L(\Psi) = Q(\Psi; \Psi^{(k)}) - E_{\Psi^{(k)}} \{ \log k(\mathbf{x} | \mathbf{y}; \Psi) | \mathbf{y} \}. \quad (4.95)$$

Oakes (1999) uses (4.95) to derive a formula for the observed information matrix $\mathbf{I}(\Psi; \mathbf{y})$. On differentiating both sides of (4.95) with respect to Ψ , we obtain

$$\partial \log L(\Psi) / \partial \Psi = \partial Q(\Psi; \Psi^{(k)}) / \partial \Psi - E_{\Psi^{(k)}} \{ \partial \log k(\mathbf{x} | \mathbf{y}; \Psi) / \partial \Psi | \mathbf{y} \}. \quad (4.96)$$

On evaluating the right-hand side of (4.96) at $\Psi^{(k)} = \Psi$, we obtain

$$\partial \log L(\Psi) / \partial \Psi = [\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi^{(k)}=\Psi}, \quad (4.97)$$

since

$$E_{\Psi} \{ \partial \log k(\mathbf{x} | \mathbf{y}; \Psi) / \partial \Psi | \mathbf{y} \} = \mathbf{0}.$$

The result (4.97) was established in Section 3.4.1; see equation (3.15).

On differentiating both sides of 4.97) with respect to Ψ and $\Psi^{(k)}$, respectively, we have

$$\begin{aligned} \partial^2 \log L(\Psi) / \partial \Psi \partial \Psi^T &= \partial Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T \\ &\quad - E_{\Psi^{(k)}} \{ \partial \log k(\mathbf{x} | \mathbf{y}; \Psi) / \partial \Psi \partial \Psi^T | \mathbf{y} \} \end{aligned} \quad (4.98)$$

and

$$\begin{aligned} \mathbf{O} &= \partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^{(k)^T} \\ &\quad - E_{\Psi^{(k)}} \{ \{\partial \log k(\mathbf{x} | \mathbf{y}; \Psi) / \partial \Psi\} E_{\Psi^{(k)}} \{ \partial \log k(\mathbf{x} | \mathbf{y}; \Psi) / \partial \Psi^{(k)^T} \} \}. \end{aligned} \quad (4.99)$$

On putting $\Psi^{(k)} = \Psi$ in the right-hand side of (4.98) and (4.99) and then adding these two equations, we obtain on taking the negative of the resulting expression,

$$\begin{aligned} -\partial^2 \log L(\Psi) / \partial \Psi \partial \Psi^T &= -[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T \\ &\quad + \partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^{(k)^T}]_{\Psi^{(k)}=\Psi}, \end{aligned} \quad (4.100)$$

since

$$\begin{aligned} E_{\Psi} \{ \partial^2 \log k(\mathbf{x} | \mathbf{y}; \Psi) / \partial \Psi \partial \Psi^T | \mathbf{y} \} \\ = -E_{\Psi} \{ \{\partial \log k(\mathbf{x} | \mathbf{y}; \Psi) / \partial \Psi\} E_{\Psi} \{ \partial \log k(\mathbf{x} | \mathbf{y}; \Psi) / \partial \Psi^T \} \}. \end{aligned} \quad (4.101)$$

By evaluating the right-hand side of (4.100) at $\Psi = \hat{\Psi}$, we get the observed information matrix $\mathbf{I}(\hat{\Psi}; \mathbf{y})$. This is Oakes' formula.

In the case where the complete data is from an exponential family with sufficient statistics $t(\mathbf{x})$ (see Section 1.5.3), it follows from equation (1.60) that

$$[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^{(k)^T}]_{\Psi^{(k)}=\Psi} = \partial E_{\Psi} \{ t(\mathbf{X}) | \mathbf{y} \} / \partial \Psi. \quad (4.102)$$

4.7.4 Example 4.6: Oakes' Standard Error for Example 1.1

Let us compute standard error for Example 1.1 of Section 1.4.2 using Oakes' formula. From equation (1.22), we get

$$Q(\Psi; \Psi^{(k)}) = \frac{\Psi^{(k)} y_1}{2 + \Psi^{(k)}} \log \Psi + (y_2 + y_3) \log(1 - \Psi) + y_4 \log \Psi \quad (4.103)$$

leading to

$$-[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi^{(k)} = \Psi} = \frac{y_1}{\Psi(2 + \Psi)} + \frac{y_2 + y_3}{(1 - \Psi)^2} + \frac{y_4}{\Psi^2} \quad (4.104)$$

and

$$-[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^{(k)}]_{\Psi^{(k)} = \Psi} = -\frac{2y_1}{\Psi(2 + \Psi)^2}. \quad (4.105)$$

Adding equations (4.104) and (4.105), we get (1.16), which was obtained directly. This verifies Oakes' formula for this example.

4.7.5 Example 4.7: Louis' Method for Example 2.4

Let us first illustrate Louis' method for our Example 2.4. We start with the complete-data log likelihood $\log L_c(\Psi)$ in (2.25) with expressions for n_A^+, n_B^+ , and n_O^+ as

$$\begin{aligned} n_A^+ &= n_{AA} + \frac{1}{2}n_{AO} + \frac{1}{2}n_{AB} = n_A - \frac{1}{2}n_{AO} + \frac{1}{2}n_{AB}, \\ n_B^+ &= n_{BB} + \frac{1}{2}n_{BO} + \frac{1}{2}n_{AB} = n_B - \frac{1}{2}n_{BO} + \frac{1}{2}n_{AB}, \end{aligned}$$

and

$$n_O^+ = n_O + \frac{1}{2}n_{AO} + \frac{1}{2}n_{BO}.$$

Let us denote the conditional expectations of n_A^+, n_B^+, n_O^+ given the observed data \mathbf{y} , using $\Psi = \hat{\Psi}$, by n_A^*, n_B^*, n_O^* respectively; these are easily obtained using the conditional binomial distributions of n_{AO} and n_{BO} . Now on taking the conditional expectation of $\log L_c(\Psi)$ given \mathbf{y} with $\Psi = \hat{\Psi}$, we have that

$$Q(\Psi; \hat{\Psi}) = 2\{n_A^* \log(p) + n_B^* \log(q) + n_O^* \log(r)\},$$

from which we obtain

$$\partial Q(\Psi; \hat{\Psi}) / \partial \Psi = 2 \left(\begin{array}{c} \frac{n_A^*}{p} - \frac{n_O^*}{r} \\ \frac{n_B^*}{q} - \frac{n_O^*}{r} \end{array} \right)$$

and

$$-[\partial^2 Q(\Psi; \hat{\Psi}) / \partial \Psi \partial \Psi^T]_{\Psi = \hat{\Psi}} = 2 \left(\begin{array}{cc} \frac{1}{\hat{p}^2} n_A^* + \frac{1}{\hat{r}^2} n_O^* & \frac{1}{\hat{r}^2} n_O^* \\ \frac{1}{\hat{q}^2} n_B^* + \frac{1}{\hat{r}^2} n_O^* & + \frac{1}{\hat{r}^2} n_O^* \end{array} \right). \quad (4.106)$$

Using the fact that n_{AO} and n_{BO} have independent binomial distributions, we find that the $\text{cov}_{\Psi} \{S_c(\mathbf{X}; \Psi) | \mathbf{y}\}$ is given by

$$2 \left(\begin{array}{cc} n_A \frac{(p+r)^2}{pr(p+2r)^2} + n_B \frac{q}{r(q+2r)^2} & n_A \frac{(p+r)}{4r(p+2r)^2} + n_O \frac{q+r}{4r(q+2r)^2} \\ n_A \frac{(p+r)}{4r(p+2r)^2} + n_O \frac{q+r}{4r(q+2r)^2} & n_B \frac{(q+r)^2}{qr(q+2r)^2} + n_A \frac{p}{r(p+2r)^2} \end{array} \right).$$

On subtracting this last expression evaluated at $\Psi = \hat{\Psi}$ from (4.106), we obtain the observed information matrix $I(\hat{\Psi}; \mathbf{y})$. We computed this difference for our data and inverted it to obtain an estimate of the covariance matrix of the MLE of $\Psi = (p, q)^T$,

$$I^{-1}(\hat{\Psi}; \mathbf{y}) = \begin{pmatrix} 0.000788531 & -0.000287210 \\ -0.000287210 & 0.000672423 \end{pmatrix}. \quad (4.107)$$

It can be seen for this example that the inverse of the observed information matrix $I^{-1}(\hat{\Psi}; \mathbf{y})$ is rather different to the inverse of the estimated expected information matrix, $\mathcal{I}_c(\hat{\Psi})$, given by (2.23) in Section 2.4.

4.7.6 Baker's Method for Standard Error for Categorical Data

We shall describe briefly Baker's method for computing the observed information matrix in the case of categorical data and apply it to the example of Section 2.4; for details see Baker (1992). We need some notations for matrix and vector operations.

Here A and B are matrices and v is a vector. Operations mentioned here are evidently defined only when the orders of the matrices conform appropriately:

$A \cdot B$: element-by-element multiplication of A and B

A/B : element-by-element division of A by B

$\mathbf{1}$: a column vector of 1's

$\text{diag}(v)$: diagonal matrix with elements of v in the diagonal

$\text{block}(A, B)$: a matrix with A and B forming a block diagonal, with 0 matrices elsewhere

$\text{hcat}(A, B)$: horizontal concatenation of A and B

$\text{vcat}(A, B)$: vertical concatenation of A and B .

The last three associative operations can have more than two arguments and can be applied to indexed arguments, as in $\text{block}_{|\kappa} \text{matrix}(\kappa)$. The categorical data are assumed to follow the Poisson, multinomial or product-multinomial model with parameter d -vector Ψ . Let $\mathbf{y} = (y_1, \dots, y_M)^T$ be the cell counts in the incomplete data and $\mathbf{x} = (x_1, \dots, x_N)^T$ be the cell counts in the complete data with $N \geq M$. Let $\mathbf{U}(\Psi)$ and $\mathbf{V}(\Psi)$ denote $E(\mathbf{Y})$ and $E(\mathbf{X})$, respectively, where \mathbf{U} and \mathbf{V} are related by $\mathbf{U}(\Psi) = \mathbf{C}\mathbf{V}(\Psi)$.

Write $\mathbf{V}(\Psi)$ as

$$\mathbf{V}(\Psi) \propto \exp\left[\sum_{\kappa=1}^K \mathbf{G}^{(\kappa)} \mathbf{T}^{(\kappa)}\right],$$

where $\mathbf{T}^{(\kappa)} = t^{(\kappa)}(\mathbf{X}^{(\kappa)} \Psi^{(\kappa)}; \mathbf{Z}^{(\kappa)})$ is a m -vector, $\Psi^{(\kappa)}$ is a $d^{(\kappa)}$ -vector of a subset of the parameters with $\sum_{\kappa=1}^K d^{(\kappa)} = d$, $\mathbf{X}^{(\kappa)}$ is a $m \times d^{(\kappa)}$ design matrix, $\mathbf{Z}^{(\kappa)}$ is a m -vector, and $\mathbf{G}^{(\kappa)}$ is an $N \times m$ matrix. The function $t^{(\kappa)}$ operates on each element of its vector arguments; see Baker (1992, Page 69).

Let us denote by $\mathbf{T}'^{(\kappa)}$ and $\mathbf{T}''^{(\kappa)}$ the vector of the first and second derivatives, respectively, of the components of $\mathbf{T}^{(\kappa)}$ with respect to the corresponding elements of $(\mathbf{X}_q^{(\kappa)} \Psi^{(\kappa)})$. Let

$$\mathbf{R} = \mathbf{1} - \mathbf{C}^T (\mathbf{Y}/\mathbf{U}),$$

$$\mathbf{S} = \text{hcat}_{|\kappa} \mathbf{G}^{(\kappa)} \text{diag} \mathbf{T}'^{(\kappa)} \mathbf{X}^{(\kappa)},$$

$$\mathbf{I}_1 = \text{block}_{|\kappa} \mathbf{X}^{(\kappa)^T} \text{diag} \mathbf{T}''^{(\kappa)} \cdot (\mathbf{G}^{(\kappa)^T} (\mathbf{R} \cdot \mathbf{V})) \mathbf{X}^{(\kappa)},$$

$$\mathbf{I}_2 = \mathbf{S}^T \text{diag}(\mathbf{R} \cdot \mathbf{V}) \mathbf{S},$$

$$\mathbf{I}_3 = \mathbf{S}^T \text{diag}(\mathbf{V}) \mathbf{C}^T \text{diag}(\mathbf{Y}/(\mathbf{U} \cdot \mathbf{U})) \mathbf{C} \text{diag}(\mathbf{V}) \mathbf{S}.$$

Then the observed information matrix \mathbf{I} is obtained as

$$\mathbf{I}(\hat{\Psi}; \mathbf{y}) = \mathbf{I}_1 + \mathbf{I}_2 + \mathbf{I}_3. \quad (4.108)$$

The expected information matrix is obtained as

$$\mathcal{I}(\hat{\Psi}) = \mathbf{S}^T \text{diag}(\mathbf{V}) \mathbf{C}^T \text{diag}(1/\mathbf{Y}) \mathbf{C} \text{diag}(\mathbf{V}) \mathbf{S}. \quad (4.109)$$

4.7.7 Example 4.8: Baker's Method for Example 2.4

In our example of Section 2.4, $N = 6$, $M = 4$, $n = 435$, $d = 2$, $\Psi = \begin{pmatrix} p \\ q \end{pmatrix}$. Note that $r = 1 - p - q$.

$$\mathbf{U}(\Psi) = N \begin{pmatrix} r^2 \\ p^2 + 2pr \\ q^2 + 2qr \\ 2pq \end{pmatrix}, \quad \mathbf{V}(\Psi) = N \begin{pmatrix} r^2 \\ p^2 \\ 2pr \\ q^2 \\ 2qr \\ 2pq \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

It can be seen that the following choices make

$$\mathbf{V} = n \exp(\mathbf{GT}),$$

$$\mathbf{X} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} e \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{T} = \log(\mathbf{Z} + \mathbf{X}\Psi),$$

$$\mathbf{G} = \begin{pmatrix} 0 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ \log(2) & 1 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ \log(2) & 0 & 1 & 1 \\ \log(2) & 1 & 1 & 0 \end{pmatrix}.$$

With our data, we obtain by inverting the observed information matrix computed from (4.108), the estimated covariance matrix of the MLE of $\Psi = (p, q)^T$, $\hat{\Psi}$, to be the same as (4.107). Using (4.109) to calculate the estimated expected information matrix, we obtain the same estimate of the covariance matrix of $\hat{\Psi}$ as given by (2.23) in Section 2.4. There the expected information matrix $\mathcal{I}(\Psi)$ was calculated by working directly with the incomplete-data log likelihood function $\log L(\Psi)$.

Kim and Taylor (1995) also described ways of estimating the covariance matrix using computations that are a part of the EM steps.

4.8 ACCELERATION OF THE EM ALGORITHM VIA AITKEN'S METHOD

4.8.1 Aitken's Acceleration Method

The most commonly used method for EM acceleration is the multivariate version of Aitken's acceleration method. Suppose that $\Psi^{(k)} \rightarrow \Psi^*$, as $k \rightarrow \infty$. Then we can express Ψ^* as

$$\Psi^* = \Psi^{(k)} + \sum_{h=1}^{\infty} (\Psi^{(h+k)} - \Psi^{(h+k-1)}). \quad (4.110)$$

Now

$$\Psi^{(h+k)} - \Psi^{(h+k-1)} = M(\Psi^{(h+k-1)}) - M(\Psi^{(h+k-2)}) \quad (4.111)$$

$$\approx J(\Psi^{(h+k-2)})(\Psi^{(h+k-1)} - \Psi^{(h+k-2)}) \quad (4.112)$$

$$\approx J(\Psi^*)(\Psi^{(h+k-1)} - \Psi^{(h+k-2)}), \quad (4.113)$$

since

$$J(\Psi^{(h+k)}) = J(\Psi^*)$$

for k sufficiently large. The approximation (4.112) to (4.111) is obtained by a linear Taylor series expansion of $M(\Psi^{(h+k-1)})$ about the point $\Psi^{(h+k-2)}$, as in (3.67). Repeated application of (4.113) in (4.110) gives

$$\begin{aligned} \Psi^* &\approx \Psi^{(k)} + \sum_{h=0}^{\infty} \{J(\Psi^*)\}^h (\Psi^{(k+1)} - \Psi^{(k)}) \\ &= \Psi^{(k)} + \{I_d - J(\Psi^*)\}^{-1} (\Psi^{(k+1)} - \Psi^{(k)}), \end{aligned} \quad (4.114)$$

as the power series

$$\sum_{h=0}^{\infty} \{J(\Psi^*)\}^h$$

converges to $\{I_d - J(\Psi^*)\}^{-1}$ if $J(\Psi^*)$ has all its eigenvalues between 0 and 1.

4.8.2 Louis' Method

The multivariate version (4.114) of Aitken's acceleration method suggests trying the sequence of iterates $\{\Psi_A^{(k)}\}$, where $\Psi_A^{(k+1)}$ is defined by

$$\Psi_A^{(k+1)} = \Psi_A^{(k)} + \{I_d - J(\Psi_A^{(k)})\}^{-1} (\Psi_{EMA}^{(k+1)} - \Psi_A^{(k)}), \quad (4.115)$$

where $\Psi_{EMA}^{(k+1)}$ is the EM iterate produced using $\Psi_A^{(k)}$ as the current fit for Ψ .

Hence this method proceeds on the $(k+1)$ th iteration by first producing $\Psi_{EMA}^{(k+1)}$ using an EM iteration with $\Psi_A^{(k)}$ as the current fit for Ψ . One then uses the EM iterate $\Psi_{EMA}^{(k+1)}$ in Aitken's acceleration procedure (4.114) to yield the final iterate $\Psi_A^{(k+1)}$ on the $(k+1)$ th iteration. This is the method proposed by Louis (1982) for speeding up the convergence of the EM algorithm.

Louis (1982) suggests making use of the relationship (3.73) to estimate $J(\Psi_A^{(k)})$ in (4.115). Using (3.73) in (4.115) gives

$$\Psi_A^{(k+1)} = \Psi_A^{(k)} + I^{-1}(\Psi_A^{(k)}; \mathbf{y}) \mathcal{I}_c(\Psi_A^{(k)}; \mathbf{y})(\Psi_{EMA}^{(k+1)} - \Psi_A^{(k)}). \quad (4.116)$$

As cautioned by Louis (1982), the relationship (3.73) is an approximation useful only local to the MLE, and so should not be used until some EM iterations have been performed. As noted by Meilijson (1989), the use of (4.116) is approximately equivalent to using the Newton-Raphson algorithm to find a zero of the (incomplete-data) score statistic $S(\mathbf{y}; \Psi)$. To see this, suppose that the EM iterate $\Psi_{EMA}^{(k+1)}$ satisfies the condition

$$[\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi_{EMA}^{(k+1)}} = \mathbf{0}.$$

Then from (3.79), we have that

$$S(\mathbf{y}; \Psi_A^{(k)}) \approx \mathcal{I}_c(\Psi_A^{(k)}; \mathbf{y})(\Psi_{EMA}^{(k+1)} - \Psi_A^{(k)}). \quad (4.117)$$

On substituting this result in (4.116), we obtain

$$\Psi_A^{(k+1)} \approx \Psi_A^{(k)} + \mathbf{I}^{-1}(\Psi_A^{(k)}; \mathbf{y})S(\mathbf{y}; \Psi_A^{(k)}). \quad (4.118)$$

It can be seen from (1.6) that the right-hand side of (4.118) is the iterate produced on the $(k+1)$ th iteration of the Newton-Raphson procedure applied to finding a zero of $S(\mathbf{y}; \Psi)$. Hence the use of Aitken's procedure as applied by Louis (1982) is essentially equivalent to the Newton-Raphson procedure applied to $S(\mathbf{y}; \Psi)$.

Meilijson (1989) and Jamshidian and Jennrich (1993) note that the accelerated sequence (4.115) as proposed by Louis (1982) is precisely the same as that obtained by applying the Newton-Raphson method to find a zero of the difference

$$\delta(\Psi) = \mathbf{M}(\Psi) - \Psi,$$

where \mathbf{M} is the map defined by the EM sequence. To see this, we can write (4.115) as

$$\Psi_A^{(k+1)} = \Psi_A^{(k)} + \{\mathbf{I}_d - \mathbf{J}(\Psi_A^{(k)})\}^{-1}\{\mathbf{M}(\Psi_A^{(k)}) - \Psi_A^{(k)}\}, \quad (4.119)$$

since

$$\Psi_{EMA}^{(k+1)} = \mathbf{M}(\Psi_A^{(k)}).$$

As the gradient of $\delta(\Psi)$ is $-\{\mathbf{I}_d - \mathbf{J}(\Psi)\}$, (4.119) corresponds exactly to applying the Newton-Raphson procedure to find a zero of $\delta(\Psi)$.

On further ways to approximate $\{\mathbf{I}_d - \mathbf{J}(\Psi_A^{(k)})\}$ for use in (4.119), Meilijson (1989) suggests using symmetric quasi-Newton updates. However, as cautioned by Jamshidian and Jennrich (1993), these will not work, as $\{\mathbf{I}_d - \mathbf{J}(\Psi)\}$ is in general not symmetric.

4.8.3 Example 4.9: Multinomial Data

As a simple illustration of the use of Aitken's acceleration procedure as proposed by Louis (1982), we consider an example from his paper in which he applies his formula (4.118) to the multinomial data in Example 1.1. It was applied after two consecutive EM iterations. In this case,

$$\begin{aligned} \Psi_A^{(2)} &= \Psi^{(2)} \\ &= 0.626338, \end{aligned}$$

and

$$\begin{aligned} \mathbf{I}^{-1}(\Psi_A^{(2)}; \mathbf{y})\mathcal{I}_c(\Psi_A^{(2)}; \mathbf{y}) &= (434.79/376.95) \\ &= 1.153442. \end{aligned}$$

Thus

$$\begin{aligned}\Psi_A^{(3)} &= 0.626338 + 1.153442(\Psi_{EMA}^{(3)} - 0.626338) \\ &= 0.6268216,\end{aligned}$$

since $\Psi_{EMA}^{(3)} = 0.626757$. It can be seen that $\Psi_A^{(3)}$ is closer to $\hat{\Psi} = 0.6268215$ than $\Psi^{(4)}$ (0.626812), the fourth iterate with the original (unaccelerated) EM algorithm.

4.8.4 Example 4.10: Geometric Mixture

We report here an example from Meilijson (1989), who considers the application of the EM algorithm to ML estimation of the parameter vector

$$\boldsymbol{\Psi} = (\pi_1, p_1, p_2)^T$$

in the two-component geometric mixture

$$f(w; \boldsymbol{\Psi}) = \sum_{i=1}^2 \pi_i f(w; p_i),$$

where

$$f(w; p_i) = p_i(1 - p_i)^{w-1}, \quad w = 1, 2, \dots \quad (0 \leq p_i \leq 1),$$

for $i = 1, 2$.

Proceeding as in Section 1.4 for the general case of a finite mixture model, we declare the complete-data vector \mathbf{x} as

$$\mathbf{x} = (\mathbf{y}^T, \mathbf{z}^T)^T,$$

where $\mathbf{y} = (w_1, \dots, w_n)^T$ contains the observed data and where the missing data vector \mathbf{z} is taken to be

$$\mathbf{z} = (z_1^T, \dots, z_n^T)^T.$$

As in Section 1.4, the $z_{ij} = (z_j)_i$ is taken to be one or zero, according as to whether the j th observation arises or does not arise from the i th component of the mixture ($i = 1, 2$; $j = 1, \dots, n$).

The complete-data log likelihood is given by

$$\begin{aligned}\log L_c(\boldsymbol{\Psi}) &= \sum_{i=1}^2 \sum_{j=1}^n z_{ij} \{ \log \pi_i + \log p_i \\ &\quad + (w_j - 1) \log(1 - p_i) \} \quad (4.120)\end{aligned}$$

$$\begin{aligned}&= \sum_{i=1}^2 \{ n_i (\log \pi_i + \log p_i) \\ &\quad + (\sum_{j=1}^n z_{ij} w_j - n_i) \log(1 - p_i) \}, \quad (4.121)\end{aligned}$$

where

$$n_i = \sum_{j=1}^n z_{ij}.$$

Corresponding to the E-step at the $(k+1)$ th iteration of the EM algorithm, the Q -function is given by

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) &= \sum_{i=1}^2 \{n_i^{(k)} (\log \pi_i + \log p_i) \\ &\quad + (\sum_{j=1}^n z_{ij}^{(k)} w_j - n_i^{(k)}) \log(1 - p_i)\}, \end{aligned} \quad (4.122)$$

where

$$n_i^{(k)} = \sum_{j=1}^n z_{ij}^{(k)}$$

and

$$z_{ij}^{(k)} = \frac{\pi_i^{(k)} p_i^{(k)} (1 - p_i^{(k)})^{w_j-1}}{\sum_{h=1}^2 \pi_h^{(k)} p_h^{(k)} (1 - p_h^{(k)})^{w_j-1}}$$

is the current conditional expectation of Z_{ij} given \mathbf{y} (that is, the current posterior probability that the j th observation arises from the i th component of the mixture ($i = 1, 2$)).

As in the general case (1.37) of a finite mixture model,

$$\pi_i^{(k+1)} = n_i^{(k)} / n,$$

while specific to geometric components in the mixture, we have on differentiation of (4.122) that

$$p_i^{(k+1)} = (\sum_{j=1}^n z_{ij}^{(k)} w_j / n_i^{(k)})^{-1} \quad (i = 1, 2).$$

We now consider the calculation of the empirical information matrix $I_e(\Psi; \mathbf{y})$ as defined by (4.41) after the k th iteration. It can be seen from this definition, that we have to calculate $s(w_j)$, the (incomplete-data) score statistic based on just the j th observation for each j ($j = 1, \dots, n$). It can be expressed in terms of the complete-data single-observation score statistics as

$$\begin{aligned} s(w_j; \Psi) &= \partial \log L_j(\Psi) / \partial \Psi \\ &= E_{\Psi^{(k)}} \{ \partial \log L_{cj}(\Psi) / \partial \Psi \mid \mathbf{y} \} \\ &= \partial Q_j(\Psi; \Psi^{(k)}) / \partial \Psi, \end{aligned}$$

where

$$Q_j(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log L_{cj}(\Psi) \mid \mathbf{y} \}.$$

From (4.121),

$$\log L_{cj}(\Psi) = \sum_{i=1}^2 z_{ij} \{ \log \pi_i + \log p_i + (w_j - 1) \log(1 - p_i) \},$$

and so

$$Q_j(\Psi; \Psi^{(k)}) = \sum_{i=1}^2 z_{ij}^{(k)} \{ \log \pi_i + \log p_i + (w_j - 1) \log(1 - p_i) \}. \quad (4.123)$$

We let $s_i(w_j; \Psi)$ denote the i th element of $\mathbf{s}(w_j; \Psi)$ for $i = 1, 2$, and 3 , corresponding to the parameters π_1 , p_1 , and p_2 . On differentiation of (4.123) with respect to Ψ , we have that

$$s_1(w_j; \Psi) = (z_{1j}^{(k)} - \pi_1)/(\pi_1\pi_2) \quad (4.124)$$

and

$$s_{i+1}(w_j; \Psi) = z_{ij}^{(k)}(1 - w_j p_i)/\{p_i(1 - p_i)\} \quad (i = 1, 2). \quad (4.125)$$

The empirical covariance matrix $\mathbf{I}_e(\Psi; \mathbf{y})$ is therefore given by (4.41) with $\mathbf{s}(w_j; \Psi)$ specified by (4.124) and (4.125).

Meilijson (1989) tries various methods to fit the geometric mixture model by maximum likelihood to a data set involving approximately 5,000 observations with values of W ranging from 1 to 14. The MLE of Ψ is

$$\hat{\Psi} = (0.40647, 0.35466, 0.90334)^T.$$

The initial value of Ψ used for the EM algorithm and its progress on the 5th, 10th, and 100th iterations are displayed in Table 4.3.

Table 4.3 Results of the EM Algorithm for the Geometric Mixture Model.

Iteration (k)	$\pi_1^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$
0	0.2	0.01	0.2
5	0.146	0.248	0.704
10	0.201	0.264	0.767
100	0.405	0.354	0.902

Meilijson (1989) finds that the Newton-Raphson and Louis methods, which are practically the same, converged after five iterations (monotonically and very smoothly), starting from the EM iterate $\Psi^{(5)}$. The Newton-Raphson method modified by using the empirical information matrix $\mathbf{I}_e(\Psi^{(k)}; \mathbf{y})$ in place of the observed information matrix $\mathbf{I}(\Psi^{(k)}; \mathbf{y})$ on each iteration k , converged after eight iterations from $\Psi^{(5)}$, the second of which was an unstable overshoot, as can be seen from Table 4.4.

The data set was modified by Meilijson (1989) to fit the geometric mixture model less perfectly. For this modified set, the MLE was

$$\hat{\Psi} = (0.64, 0.35, 0.86)^T.$$

The Newton-Raphson method with the modification of the empirical covariance matrix converged in six iterations from

$$\Psi = (0.4, 0.8, 0.9)^T,$$

but diverged from

$$\Psi = (0.2, 0.5, 0.7)^T.$$

However, it did converge in eight iterations from the latter point when negative probabilities were interpreted as 0.01 and those exceeding unity as 0.99. The Newton-Raphson and Louis methods were found to be fast and generally smooth, but with a smaller radius of convergence than under the perfect fit.

Table 4.4 Results of Newton-Raphson Method using Empirical Information Matrix for Geometric Mixture Model.

Iteration (k)	$\pi_1^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$
0	0.146	0.248	0.704
1	0.43	0.52	0.87
2	0.60	0.49	0.986
3	0.45	0.42	0.91
4	0.42	0.37	0.91
5	0.406	0.356	0.903
.	.	.	.
.	.	.	.
8	0.406	0.355	0.903

4.8.5 Example 4.11: Grouped and Truncated Data. (*Example 2.8 Continued*)

We return now to Example 2.9 in Section 2.8.7 with the numerical illustration thereof, where the two-component log normal mixture model (2.59) was fitted to some grouped and truncated data on the volume of red blood cells of a cow. The EM algorithm was found to take 117 iterations to converge. For this data set, the largest eigenvalue λ_{\max} of the rate matrix (3.72) is 0.975. Values of λ_{\max} near one indicate a slow rate of convergence. Jones and McLachlan (1992) investigated various ways of speeding up the EM algorithm in its application to this data set. In Table 4.5, we report their results obtained using the EM algorithm and its accelerated version using Louis' method. It can be seen that the latter is effective here in reducing the number of iterations.

They also tried the modified Newton-Raphson method with the information matrix replaced by the empirical information matrix for grouped data, as given by (4.45). It gave identical results for this data set. This is not unexpected, as we have noted that Louis' method is essentially the same as the Newton-Raphson method. But Louis' method and the modified Newton-Raphson method will not always give the same results, as demonstrated by Jones and McLachlan (1992) in their application of these two methods to another data set of this type, where Louis' method was not as effective as the modified Newton-Raphson method in terms of number of iterations. The standard errors of the estimates, given in parentheses in Table 4.5, are based on using the inverse of the empirical information matrix (4.45) to approximate the covariance matrix of the MLE.

4.9 AN AITKEN ACCELERATION-BASED STOPPING CRITERION

The stopping criterion usually adopted with the EM algorithm is in terms of either the size of the relative change in the parameter estimates or the log likelihood. As Lindstrom and Bates (1988) emphasize, this is a measure of lack of progress but not of actual convergence. Böhning, Dietz, Schaub, Schlattmann, and Lindsay (1994) exploit Aitken's acceleration procedure in its application to the sequence of log likelihood values to provide a useful estimate of its limiting value. It is applicable in the case where the sequence of log likelihood

Table 4.5 Comparison of Methods to Speed up the EM Algorithm for Red Blood Cell Volume Data.

Parameter	Initial Value	EM	Louis' Method
π_1	0.45	0.4521 (0.0521)	0.4530 (0.0522)
μ_1	4.00	4.0728 (0.0384)	4.0734 (0.0384)
μ_2	4.60	4.7165 (0.0242)	4.7169 (0.0242)
σ_1^2	0.08	0.0575 (0.0107)	0.0577 (0.0107)
σ_2^2	0.05	0.0438 (0.0099)	0.0436 (0.0099)
No. of iterations		107	11

Source: Adapted from Jones and McLachlan (1992), with permission of the Journal of Statistical Computation and Simulation.

values $\{l^{(k)}\}$ is linearly convergent to some value l^* , where here for brevity of notation

$$l^{(k)} = \log L(\Psi^{(k)}).$$

Under this assumption,

$$l^{(k+1)} - l^* \approx c(l^{(k)} - l^*), \quad (4.126)$$

for all k and some c ($0 < c < 1$). The equation (4.126) can be rearranged to give

$$l^{(k+1)} - l^{(k)} \approx (1 - c)(l^* - l^{(k)}), \quad (4.127)$$

for all k . It can be seen from (4.127) that, if c is very close to one, a small increment in the log likelihood, $l^{(k+1)} - l^{(k)}$, does not necessarily mean that $l^{(k)}$ is very close to l^* .

From (4.127), we have that

$$l^{(k+1)} - l^{(k)} \approx c(l^{(k)} - l^{(k-1)}), \quad (4.128)$$

for all k , corresponding to the (multivariate) expression (4.113) for successive increments in the estimates of the parameter vector. Just as Aitken's acceleration procedure was applied to (4.110) to obtain (4.114), we can apply it to (4.128) to obtain the corresponding result for the limit l^* of the sequence of log likelihood values

$$l^* = l^{(k)} + \frac{1}{(1 - c)}(l^{(k+1)} - l^{(k)}). \quad (4.129)$$

Since c is unknown, it has to be estimated in (4.129), for example, by the ratio of successive increments,

$$c^{(k)} = (l^{(k+1)} - l^{(k)})/(l^{(k)} - l^{(k-1)}).$$

This leads to the Aitken accelerated estimate of l^* ,

$$l_A^{(k+1)} = l^{(k)} + \frac{1}{(1 - c^{(k)})} (l^{(k+1)} - l^{(k)}). \quad (4.130)$$

In applications where the primary interest is in the sequence of log likelihood values rather than the sequence of parameter estimates, Böhning et al. (1994) suggest the EM algorithm can be stopped if

$$|l_A^{(k+1)} - l_A^{(k)}| < \text{tol},$$

where tol is the desired tolerance. An example concerns the resampling approach (McLachlan, 1987) to the problem of assessing the null distribution of the likelihood ratio test statistic for the number of components in a mixture model. The criterion (4.130) is applicable for any log likelihood sequence that is linearly convergent.

4.10 CONJUGATE GRADIENT ACCELERATION OF EM ALGORITHM

4.10.1 Conjugate Gradient Method

Jamshidian and Jennrich (1993) propose an alternative method based on conjugate gradients. The key is that the change in Ψ after an EM iteration

$$\delta(\Psi) = M(\Psi) - \Psi$$

can be viewed (approximately at least) as a generalized gradient, making it natural to apply generalized conjugate gradient methods in an attempt to accelerate the EM. Though not identified as such, this was done by Golub and Nash (1982) as their alternative to the Yates (1933) EM algorithm for fitting unbalanced analysis of variance models. The proposed method is relatively simple and can handle problems with a large number of parameters. The latter is an area where the EM algorithm is particularly important and where it is often the only algorithm used. Before we discuss the generalized gradient approach to accelerating the performance of the EM algorithm, we describe the generalized conjugate gradient algorithm used by Jamshidian and Jennrich (1993).

4.10.2 A Generalized Conjugate Gradient Algorithm

We present the algorithm in Jamshidian and Jennrich (1993) for finding the maximum of a function $h(\Psi)$, where Ψ ranges over a subset of \mathbb{R}^d . Let $q(\Psi)$ denote the gradient of $h(\Psi)$ and consider the generalized norm

$$\|\Psi\| = (\Psi^T A \Psi)^{1/2}$$

on \mathbb{R}^d , where A is a positive definite matrix. Also, let $\tilde{q}(\Psi)$ be the gradient of $h(\Psi)$ with respect to this norm. That is,

$$\tilde{q}(\Psi) = A^{-1} q(\Psi).$$

The vector $\tilde{q}(\Psi)$ is called the generalized gradient of $h(\Psi)$ defined by A .

The generalized conjugate gradient algorithm is implemented as follows. Let $\Psi^{(0)}$ be the initial value of Ψ and let

$$d^{(0)} = \tilde{q}(\Psi^{(0)}).$$

One sequentially computes for $k = 0, 1, 2, \dots$,

1. $\alpha^{(k)}$, the value of α that maximizes

$$h(\Psi^{(k)} + \alpha d^{(k)});$$

2. $\Psi^{(k+1)} = \Psi^{(k)} + \alpha^{(k)} d^{(k)}$;
3. $\beta^{(k)} = \frac{\{q^T(\Psi^{(k+1)}) - q^T(\Psi^{(k)})\}\tilde{q}(\Psi^{(k+1)})}{\{q^T(\Psi^{(k+1)}) - q^T(\Psi^{(k)})\}d^{(k)}}$;
4. $d^{(k+1)} = \tilde{q}(\Psi^{(k+1)}) - \beta^{(k)} d^{(k)}$.

This is called a generalized conjugate gradient algorithm, because it uses generalized gradients to define the search direction $d^{(k)}$ and because, for negative definite quadratic functions h , the $d^{(k)}$ are orthogonal in the metric defined by the negative of the Hessian of h .

4.10.3 Accelerating the EM Algorithm

From (4.117),

$$\begin{aligned} \Psi^{(k+1)} - \Psi^{(k)} &= M(\Psi^{(k)}) - \Psi^{(k)} \\ &\approx \mathcal{I}_c^{-1}(\Psi^{(k)}; \mathbf{y})S(\mathbf{y}; \Psi^{(k)}). \end{aligned}$$

Thus the change in Ψ after an EM iteration,

$$\delta(\Psi) = M(\Psi) - \Psi,$$

is approximately equal to

$$\delta(\Psi) \approx \mathcal{I}_c^{-1}(\Psi; \mathbf{y})S(\mathbf{y}; \Psi), \quad (4.131)$$

where $S(\mathbf{y}; \Psi)$ is the gradient of the log likelihood function $\log L(\Psi)$. Typically, $\mathcal{I}_c(\Psi; \mathbf{y})$ is positive definite. Thus (4.131) shows that when $M(\Psi)$ is near to Ψ , the difference $\delta(\Psi) = M(\Psi) - \Psi$ is, to a good approximation, a generalized gradient of $\log L(\Psi)$.

Jamshidian and Jennrich (1993) propose accelerating the EM algorithm by applying the generalized conjugate gradient algorithm in the previous section with

$$h(\Psi) = \log L(\Psi),$$

and

$$q(\Psi) = S(\mathbf{y}; \Psi),$$

and where the generalized gradient $\tilde{q}(\Psi)$ is given by $\delta(\Psi)$, the change in Ψ after performing each subsequent EM iteration. They call the resulting algorithm the accelerated EM (AEM) algorithm. This algorithm is applied after a few EM iterations. More specifically, Jamshidian and Jennrich (1993) suggest running the EM algorithm until twice the difference between successive values of the log likelihood falls below one. In their examples, this usually occurred after five EM iterations.

Although the AEM algorithm is fairly simple, it is more complex than the EM algorithm itself. In addition to the EM iterations, one must compute the gradient of $\log L(\Psi)$ (that is, the score statistic $S(\mathbf{y}; \Psi)$). Frequently, however, the latter is either available from the

EM code or is obtainable with minor modification. The biggest complication is the line search required in Step 1 above. An algorithm for this search is given in the Appendix of Jamshidian and Jennrich (1993). Their experience suggests that a simple line search is sufficient. To demonstrate the effectiveness of the AEM algorithm, Jamshidian and Jennrich (1993) applied it to several problems in areas that included the estimation of a covariance matrix from incomplete multivariate normal data, confirmatory factor analysis, and repeated measures analysis. In terms of floating-point operation counts, for all of the comparative examples considered, the AEM algorithm increases the speed of the EM algorithm, in some cases by a factor of 10 or more.

4.11 HYBRID METHODS FOR FINDING THE MAXIMUM LIKELIHOOD ESTIMATE

4.11.1 Introduction

Various authors, including Redner and Walker (1984), propose a hybrid approach to the computation of the MLE that would switch from the EM algorithm after a few iterations to the Newton-Raphson or some quasi-Newton method. The idea is to use the EM algorithm initially to take advantage of its good global convergence properties and to then exploit the rapid local convergence of Newton-type methods by switching to such a method. It can be seen that the method of Louis (1982) is a hybrid algorithm of this nature, for in effect after a few initial EM iterations, it uses the Newton-Raphson method to accelerate convergence after performing each subsequent EM iteration. Of course there is no guarantee that these hybrid algorithms increase the likelihood $L(\Psi)$ monotonically. Hybrid algorithms have been considered also by Atkinson (1992), Heckman and Singer (1984), Jones and McLachlan (1992), and Aitkin and Aitkin (1996).

4.11.2 Combined EM and Modified Newton-Raphson Algorithm

We now consider the work of Aitkin and Aitkin (1996) on a hybrid method that combines the EM algorithm with a modified Newton-Raphson method whereby the information matrix is replaced by the empirical information matrix. In the context of fitting finite normal mixture models, they construct a hybrid algorithm that starts with five EM iterations before switching to the modified Newton-Raphson method until convergence or until the log likelihood decreases. In the case of the latter, Aitkin and Aitkin (1996) propose halving the step size up to five times. As further step-halvings would generally leave the step size smaller than that of an EM step, if the log likelihood decreases after five step-halves, the algorithm of Aitkin and Aitkin (1996) returns to the previous EM iterate and runs the EM algorithm for a further five iterations, before switching back again to the modified Newton-Raphson method. Their choice of performing five EM iterations initially is based on the work of Redner and Walker (1984), who report that, in their experience, 95 percent of the change in the log likelihood from its initial value to its maximum generally occurs in five iterations.

Aitkin and Aitkin (1996) replicate part of the study by Everitt (1988) of the fitting of a mixture of two normal densities with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 in proportions π_1 and π_2 ,

$$f(w; \Psi) = \pi_1\phi(w; \mu_1, \sigma_1^2) + \pi_2\phi(w; \mu_2, \sigma_2^2). \quad (4.132)$$

They simulated ten random samples of size $n = 50$ from this normal mixture in each of three cases with the parameter vector

$$\Psi = (\pi_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T$$

specified as in Table 4.6. These cases are increasingly difficult to fit. The normal mixture model (4.132) is fitted to each simulated random sample using the starting values, as specified in Table 4.7. The stopping criterion of Aitkin and Aitkin (1996) is a difference in successive values of the log likelihood of 10^{-5} .

Table 4.6 Parameter Values Used in the Simulation of Two-Component Normal Mixture Models.

Mixture	p	μ_1	μ_2	σ_1^2	σ_2^2
I	0.4	0	3	0.5	1
II	0.4	0	3	1	2
III	0.2	0	3	1	2

Table 4.7 Choice of Starting Values in the Fitting of Normal Mixture Models to Simulated Data.

Mixture	p	μ_1	μ_2	σ_1^2	σ_2^2
I	0.4	0	3	0.5	1
	0.4	0	1	0.5	0.5
	0.2	1	2	1	0.5
II	0.4	0	3	1	2
	0.4	0	1	1	1
	0.2	1	2	0.5	1
III	0.2	0	3	1	2
	0.2	0	1	1	1
	0.4	1	2	0.5	1

Aitkin and Aitkin (1996) find that their hybrid algorithm required 70 percent of the time required for the EM algorithm to converge, consistently over all starting values of Ψ . They find that the EM algorithm is impressively stable. Their hybrid algorithm almost always decreases the log likelihood when the switch to the modified Newton-Raphson is first applied, and sometimes requires a large number of EM controlling steps (after full step-halving) before finally increasing the log likelihood, and then usually converging rapidly to the same maximizer as with the EM algorithm. Local maxima are encountered by both algorithms about equally often.

As is well known, mixture likelihoods for small sample sizes are badly behaved with multiple maxima. Aitkin and Aitkin (1996) liken the maximization of the normal mixture log likelihood in their simulation studies as to the progress of a “traveler following the narrow EM path up a hazardous mountain with chasms on all sides. When in sight of the summit, the modified Newton-Raphson method path leapt to the top, but when followed earlier, it caused repeated falls into the chasms, from which the traveler had to be pulled back onto the EM track”.

4.12 A GENERALIZED EM ALGORITHM BASED ON ONE NEWTON-RAPHSON STEP

4.12.1 Derivation of a Condition to be a Generalized EM Sequence

In Section 1.5 in formulating the GEM algorithm, we considered as an example of a GEM algorithm the sequence of iterates $\{\Psi^{(k)}\}$, where $\Psi^{(k+1)}$ is defined to be

$$\Psi^{(k+1)} = \Psi^{(k)} + a^{(k)} \delta^{(k)}, \quad (4.133)$$

where

$$\delta^{(k)} = -[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}}^{-1} [\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi^{(k)}}, \quad (4.134)$$

and where $0 < a^{(k)} \leq 1$. The constant $a^{(k)}$ is chosen so that

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) \geq Q(\Psi^{(k)}; \Psi^{(k)}). \quad (4.135)$$

holds; that is, so that (4.133) defines a GEM algorithm. In the case of $a^{(k)} = 1$, (4.133) defines the first iterate obtained when using the Newton-Raphson procedure to obtain a root of the equation

$$\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi = \mathbf{0}$$

on the M-step where the intent is to maximize the Q -function.

We now derive the result (1.65) that

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) - Q(\Psi^{(k)}; \Psi^{(k)}) = a^{(k)} \mathbf{S}(\mathbf{y}; \Psi^{(k)})^T \mathbf{A}^{(k)} \mathbf{S}(\mathbf{y}; \Psi^{(k)}), \quad (4.136)$$

where

$$\mathbf{A}^{(k)} = \mathcal{I}_c^{-1}(\Psi^{(k)}; \mathbf{y}) \{ \mathbf{I}_d - \frac{1}{2} a^{(k)} \tilde{\mathcal{I}}_c^{(k)}(\mathbf{y}) \mathcal{I}_c^{-1}(\Psi^{(k)}; \mathbf{y}) \} \quad (4.137)$$

and where

$$\begin{aligned} \tilde{\mathcal{I}}_c^{(k)}(\mathbf{y}) &= -[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\tilde{\Psi}^{(k)}} \\ &= E_{\Psi^{(k)}} \{ \mathbf{I}_c(\tilde{\Psi}^{(k)}; \mathbf{X}) \mid \mathbf{y} \}, \end{aligned} \quad (4.138)$$

and $\tilde{\Psi}^{(k)}$ is a point on the line segment between $\Psi^{(k)}$ and $\Psi^{(k+1)}$.

From (3.47) and (3.75), we have that

$$[\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi^{(k)}} = \mathbf{S}(\mathbf{y}; \Psi^{(k)}) \quad (4.139)$$

and

$$[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}} = -\mathcal{I}_c(\Psi^{(k)}; \mathbf{y}). \quad (4.140)$$

On using (4.139) and (4.140), we can express $\Psi^{(k+1)}$ as

$$\Psi^{(k+1)} = \Psi^{(k)} + a^{(k)} \mathcal{I}_c^{-1}(\Psi; \mathbf{y}) \mathbf{S}(\mathbf{y}; \Psi^{(k)}). \quad (4.141)$$

Now on expanding $Q(\Psi; \Psi^{(k)}) - Q(\Psi^{(k)}; \Psi^{(k)})$ in a linear Taylor series expansion about the point $\Psi = \Psi^{(k+1)}$ and using the relationships (4.138), (4.139), and (4.140), we

obtain

$$\begin{aligned}
 & Q(\Psi^{(k+1)}; \Psi^{(k)}) - Q(\Psi^{(k)}; \Psi^{(k)}) \\
 &= (\Psi^{(k+1)} - \Psi^{(k)})^T S(\mathbf{y}; \Psi^{(k)}) \\
 &\quad + \frac{1}{2} (\Psi^{(k+1)} - \Psi^{(k)})^T \tilde{\mathcal{I}}_c^{(k)}(\mathbf{y}) (\Psi^{(k+1)} - \Psi^{(k)}).
 \end{aligned} \tag{4.142}$$

On substituting (4.141) for $\Psi^{(k+1)}$ in (4.142), we obtain the result (4.136).

Typically, $\mathcal{I}_c(\Psi^{(k)}; \mathbf{y})$ is positive definite, and so from (4.137), (4.133) yields a GEM sequence if the matrix

$$\mathbf{I}_d - \frac{1}{2} a^{(k)} \tilde{\mathcal{I}}_c^{(k)}(\mathbf{y}) \mathcal{I}_c^{-1}(\Psi^{(k)}; \mathbf{y})$$

is positive definite; that is, if $a^{(k)}$ is chosen sufficiently small, as discussed in Section 1.57. In the case that the complete-data density is a member of the regular exponential family with natural parameter Ψ , we have from (4.10) that

$$\mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) = \mathcal{I}_c(\Psi^{(k)}),$$

and so $\mathcal{I}_c(\Psi^{(k)}; \mathbf{y})$ is positive definite.

4.12.2 Simulation Experiment

Rai and Matthews (1993) perform a simulation study involving multinomial data in the context of a carcinogenicity experiment to compare the GEM algorithm based on one Newton-Raphson iteration with the EM algorithm using the limiting value of the Newton-Raphson iterations on the M-step. They find that this GEM algorithm can lead to significant computational savings. This is because it was found on average that the GEM algorithm required only a few more E-steps than the EM algorithm and, with only one iteration per M-step, it thus required significantly fewer Newton-Raphson iterations over the total number of M-steps. The results of their simulation are reported in Table 4.8.

Table 4.8 Results of Simulation Study Comparing EM and a GEM Algorithm for a Multinomial Problem.

	EM Algorithm		GEM Algorithm	
	# of E steps	# of M steps	# of E steps	# of M steps
Average	108	474	112	112
Standard error	12	49	14	14

Source: Adapted from Rai and Matthews (1993), with permission of the Biometric Society.

4.13 EM GRADIENT ALGORITHM

As noted in Section 1.5.7, Lange (1995a) considers the sequence of iterates defined by (4.133), but with $a^{(k)} = 1$. He calls this algorithm the EM gradient algorithm. He also considers a modified EM gradient algorithm that has $a^{(k)} = a$ in (4.133) as a means of

inflating the current EM gradient step by a factor a to speed up convergence. Lange (1995b) subsequently uses the EM gradient algorithm to form the basis of a quasi-Newton approach to accelerate convergence of the EM algorithm. But as pointed out by Lange (1995b), the EM gradient algorithm is an interesting algorithm in its own right. Since the Newton-Raphson method converges quickly, Lange (1995b) notes that the local properties of the EM gradient algorithm are almost identical with those of the EM algorithm.

The matrix

$$[\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}} = -\mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) \quad (4.143)$$

may not be negative definite, and consequently the EM gradient algorithm is not necessarily an ascent algorithm. In practice, the matrix (4.143) is typically negative definite. In the case of the complete-data density being a member of the regular exponential family with natural parameter Ψ , we have from (3.75) and (4.13) that

$$\begin{aligned} [\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \Psi^T]_{\Psi=\Psi^{(k)}} &= -\mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) \\ &= -\mathcal{I}_c(\Psi^{(k)}), \end{aligned} \quad (4.144)$$

and so negative definiteness is automatic. In this case, the EM gradient algorithm coincides with the ascent algorithm of Titterington (1984).

One advantage of Titterington's (1984) algorithm is that it is necessarily uphill. This means that a fractional step in the current direction will certainly lead to an increase in $L(\Psi)$. The EM gradient algorithm also will have this property if $\mathcal{I}_c(\Psi^{(k)}; \mathbf{y})$ is positive definite. In some cases, a reparameterization is needed to ensure that $\mathcal{I}_c(\Psi^{(k)}; \mathbf{y})$ is positive definite. For example, as noted above, if the complete-data density is a member of the regular exponential family with natural parameter Ψ , then

$$\mathcal{I}_c(\Psi; \mathbf{y}) = \mathcal{I}_c(\Psi), \quad (4.145)$$

and so $\mathcal{I}_c(\Psi; \mathbf{y})$ is positive definite. However, as seen in Section 4.2.4, if the natural parameter is $c(\Psi)$, then although (4.145) holds at the MLE, it will not hold in general at $\Psi = \Psi^{(k)}$. Hence we need to transform Ψ to $\theta = c(\Psi)$. The reparameterization does not affect the EM algorithm; see Lansky et al. (1992).

Lange (1995a) suggests improving the speed of convergence of the EM gradient algorithm by inflating the current EM gradient step by a factor a to give

$$\Psi^{(k+1)} = \Psi^{(k)} + a\delta^{(k)} \quad (4.146)$$

where, on using (4.139) and (4.140) in (4.134),

$$\delta^{(k)} = \mathcal{I}_c^{-1}(\Psi^{(k)}; \mathbf{y})S(\mathbf{y}; \Psi^{(k)}).$$

In Section 4.12.2, we noted that Rai and Matthews (1993) had considered a sequence of iterates of the form (4.146), but with $a = a^{(k)}$ and where $a^{(k)}$ was chosen so as to ensure that it implies a GEM sequence and thereby the monotonicity of $L(\Psi^{(k)})$.

Lange (1995b) considers the choice of a in (4.146) directly in terms of $L(\Psi)$. He shows that the modified EM gradient sequence has the desirable property of being locally monotonic when $0 < a < 2$. By a second-order Taylor series expansion of $\log L(\Psi) - \log L(\Psi^{(k)})$ about the point $\Psi = \Psi^{(k)}$, we have on evaluation at the point $\Psi = \Psi^{(k+1)}$ that

$$\log L(\Psi^{(k+1)}) - \log L(\Psi^{(k)}) = \frac{1}{2}(\Psi^{(k+1)} - \Psi^{(k)})^T C^{(k)} (\Psi^{(k+1)} - \Psi^{(k)}), \quad (4.147)$$

where

$$\mathbf{C}^{(k)} = \frac{2}{a} \mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) - \mathbf{I}(\tilde{\Psi}^{(k)}; \mathbf{y}), \quad (4.148)$$

and $\tilde{\Psi}^{(k)}$ is a point on the line segment from $\Psi^{(k)}$ to $\Psi^{(k+1)}$. Now

$$\begin{aligned} \lim_{k \rightarrow \infty} \mathbf{C}^{(k)} &= \lim_{k \rightarrow \infty} \left\{ \frac{2}{a} \mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) - \mathbf{I}(\tilde{\Psi}^{(k)}; \mathbf{y}) \right\} \\ &= (\frac{2}{a} - 1) \mathcal{I}_c(\Psi^*; \mathbf{y}) - \{ \mathbf{I}(\Psi^*; \mathbf{y}) - \mathcal{I}_c(\Psi^*; \mathbf{y}) \}, \end{aligned} \quad (4.149)$$

where it is assumed that the sequence $\{\Psi^{(k)}\}$ converges to some point Ψ^* .

Assuming that $\mathcal{I}_c(\Psi; \mathbf{y})$ is positive definite, the first term on the right-hand side of (4.149) is positive definite if $0 < a < 2$.

As seen in Section 3.2,

$$H(\Psi; \Psi^{(k)}) = Q(\Psi; \Psi^{(k)}) - \log L(\Psi)$$

has a maximum at $\Psi = \Psi^{(k)}$, and so

$$[\partial^2 H(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}}$$

is negative semidefinite.

The second term on the right-side of (4.149) is negative semidefinite, since

$$\begin{aligned} &\mathbf{I}(\Psi^*; \mathbf{y}) - \mathcal{I}_c(\Psi^*; \mathbf{y}) \\ &= [\partial^2 Q(\Psi; \Psi^*) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^*} - (\partial^2 \log L(\Psi^*) / \partial \Psi \partial \Psi^T) \\ &= [\partial^2 H(\Psi; \Psi^*) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^*} \\ &= \lim_{k \rightarrow \infty} [\partial^2 H(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}}, \end{aligned} \quad (4.150)$$

which is negative semidefinite from above. Hence if $0 < a < 2$, $\lim_{k \rightarrow \infty} \mathbf{C}^{(k)}$ is a positive definite matrix because it is expressible as the difference between a positive definite matrix and a negative semidefinite matrix.

Since the eigenvalues of a matrix are defined continuously in its entries, it follows that if $0 < a < 1$, the quadratic function (4.147) is positive for k sufficiently large and $\Psi^{(k+1)} \neq \Psi^{(k)}$.

Lange (1995b) also investigates the global convergence of the EM gradient algorithm. Although he concludes that monotonicity appears to be the rule in practice, to establish convergence in theory, monotonicity has to be enforced. Lange (1995b) elects to do this by using a line search at every EM gradient step. The rate of convergence of the EM gradient algorithm is identical with that of the EM algorithm. It is possible for the EM and EM gradient algorithms started from the same point to converge to different points.

4.14 A QUASI-NEWTON ACCELERATION OF THE EM ALGORITHM

4.14.1 The Method

We have seen above that the EM gradient algorithm approximates the M-step of the EM algorithm by using one step of the Newton-Raphson method applied to find a zero of the equation

$$\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi = \mathbf{0}. \quad (4.151)$$

That is, it uses

$$\begin{aligned}\Psi^{(k+1)} &= \Psi^{(k)} - [\partial^2 Q(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}} [\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi^{(k)}} \\ &= \Psi^{(k)} + \mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) \mathbf{S}(\mathbf{y}; \Psi^{(k)}).\end{aligned}\quad (4.152)$$

If the Newton-Raphson method is applied to find a zero of the incomplete-data likelihood equation directly, we have that

$$\Psi^{(k+1)} = \Psi^{(k)} + \mathbf{I}_c(\Psi^{(k)}; \mathbf{y}) \mathbf{S}(\mathbf{y}; \Psi^{(k)}). \quad (4.153)$$

Thus the use of the EM gradient algorithm can be viewed as using the Newton-Raphson method to find a zero of the likelihood equation, but with the approximation

$$\mathbf{I}(\Psi^{(k)}; \mathbf{y}) \approx \mathcal{I}_c(\Psi^{(k)}; \mathbf{y}). \quad (4.154)$$

The quasi-Newton acceleration procedure proposed by Lange (1995b) defines $\Psi^{(k+1)}$ to be

$$\Psi^{(k+1)} = \Psi^{(k)} + \{\mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) + \mathbf{B}^{(k)}\}^{-1} \mathbf{S}(\mathbf{y}; \Psi^{(k)}). \quad (4.155)$$

We have from (3.50) that

$$\mathbf{I}(\Psi^{(k)}; \mathbf{y}) = \mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) - \mathcal{I}_m(\Psi^{(k)}; \mathbf{y}), \quad (4.156)$$

where from (3.5) and (3.52),

$$\begin{aligned}\mathcal{I}_m(\Psi^{(k)}; \mathbf{y}) &= -E_{\Psi^{(k)}} \{[\partial^2 \log k(\mathbf{X} | \mathbf{y}; \Psi) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}} | \mathbf{y}\} \\ &= -[\partial^2 H(\Psi; \Psi^{(k)}) / \partial \Psi \partial \Psi^T]_{\Psi=\Psi^{(k)}}.\end{aligned}\quad (4.157)$$

Thus the presence of the term $\mathbf{B}^{(k)}$ in (4.155) can be viewed as an attempt to approximate the Hessian of $H(\Psi; \Psi^{(k)})$ at the point $\Psi = \Psi^{(k)}$.

Lange (1995b) takes $\mathbf{B}^{(k)}$ to be based on Davidon's (1959) symmetric, rank-one update defined by

$$\mathbf{B}^{(k)} = \mathbf{B}^{(k-1)} + c^{(k)} \mathbf{v}^{(k)} \mathbf{v}^{(k)T} \quad (4.158)$$

and where the constant $c^{(k)}$ and the vector $\mathbf{v}^{(k)}$ are specified as

$$c^{(k)} = -1/(\mathbf{v}^{(k)T} \mathbf{d}^{(k)}), \quad (4.159)$$

$$\mathbf{v}^{(k)} = \mathbf{h}^{(k)} + \mathbf{B}^{(k-1)} \mathbf{d}^{(k)}. \quad (4.160)$$

Here

$$\mathbf{d}^{(k)} = \Psi^{(k)} - \Psi^{(k-1)} \quad (4.161)$$

and

$$\mathbf{h}^{(k)} = [\partial H(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi^{(k-1)}}. \quad (4.162)$$

Lange (1995b) suggests taking $\mathbf{B}^{(0)} = \mathbf{I}_d$, which corresponds to initially performing an EM gradient step. When the inner product on the denominator on the right-hand side of (4.159) is zero or is very small relative to

$$\|\mathbf{h}^{(k)} + \mathbf{B}^{(k-1)} \mathbf{d}^{(k)}\| \|\mathbf{d}^{(k)}\|.$$

Lange (1995b) suggests omitting the update and taking $\mathbf{B}^{(k)} = \mathbf{B}^{(k-1)}$.

When the matrix

$$\mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) + \mathbf{B}^{(k)}$$

fails to be positive definite, Lange (1995b) suggests replacing it by

$$\mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) + (\frac{1}{2})^m \mathbf{B}^{(k)}, \quad (4.163)$$

where m is the smallest positive integer such that (4.163) is positive definite.

In view of the identities in Ψ and $\Psi^{(k)}$, namely that

$$\partial H(\Psi; \Psi^{(k)})/\partial \Psi + \partial \log L(\Psi)/\partial \Psi = \partial Q(\Psi; \Psi^{(k)})/\partial \Psi$$

and that

$$[\partial H(\Psi; \Psi^{(k)})/\partial \Psi]_{\Psi=\Psi^{(k)}} = \mathbf{0},$$

we can express $\mathbf{h}^{(k)}$ as

$$\begin{aligned} \mathbf{h}^{(k)} &= [\partial Q(\Psi; \Psi^{(k)})/\partial \Psi]_{\Psi=\Psi^{(k-1)}} - \partial \log L(\Psi^{(k-1)})/\partial \Psi \\ &= [\partial Q(\Psi; \Psi^{(k)})/\partial \Psi]_{\Psi=\Psi^{(k-1)}} - [\partial Q(\Psi; \Psi^{(k-1)})/\partial \Psi]_{\Psi=\Psi^{(k-1)}}. \end{aligned}$$

It can be seen that almost all the relevant quantities for the quasi-Newton acceleration of the EM algorithm can be expressed in terms of $Q(\Psi; \Psi^{(k)})$ and its derivatives. The only relevant quantity not so expressible is the likelihood $L(\Psi)$. The latter needs to be computed if the progress of the algorithm is to be monitored. If it is found that (4.155) modified according to (4.163) overshoots at any given iteration, then some form of step-decrementing can be used to give an increase in $L(\Psi)$. Lange (1995b) follows Powell's (1978) suggestion and fits a quadratic to the function in $r, \log L(\Psi(r))$, through the values $L(\Psi^{(k)})$ and $L(\Psi^{(k+1)})$ with slope

$$\{\partial \log L(\Psi^{(k)})/\partial \Psi\}^T \mathbf{d}^{(k)}$$

at $r = 0$, where

$$\Psi(r) = \Psi^{(k)} + r \mathbf{d}^{(k)}.$$

If the maximum of the quadratic occurs at r_{\max} , then one steps back with

$$r = \max\{r_{\max}, 0.1\}.$$

If this procedure still does not yield an increase in $L(\Psi)$, then it can be repeated. Lange (1995b) notes that one or two step decrements invariably give the desired increase in $L(\Psi)$. Because the algorithm moves uphill, step decrementing is bound to succeed. Lange (1995b) surmises that a step-halving strategy would be equally effective.

As Lange (1995b) points out, the early stages of the algorithm resemble a close approximation to the EM algorithm, later stages approximate Newton-Raphson, and intermediate stages make a graceful transition between the two extremes.

4.14.2 Example 4.12: Dirichlet Distribution

We now consider the example given by Lange (1995b) on ML estimation of the parameters of the Dirichlet distribution. This distribution is useful in modeling data on proportions

(Kingman, 1993). We let U_1, \dots, U_m be m independent random variables with U_i having a gamma $(\theta_i, 1)$ density function,

$$f(u; \alpha, 1) = \{u^{\theta_i-1}/\Gamma(\theta_i)\} \exp(-u) I_{[0,\infty)}(u); \quad (\theta_i > 0),$$

for $i = 1, \dots, m$.

Put

$$W_i = U_i / \sum_{h=1}^m U_h \quad (i = 1, \dots, m). \quad (4.164)$$

Then the random vector

$$\mathbf{W} = (W_1, \dots, W_m)^T$$

has the Dirichlet density

$$\frac{\Gamma(\sum_{i=1}^m \theta_i)}{\prod_{i=1}^m \Gamma(\theta_i)} \prod_{i=1}^m w_i^{\theta_i-1} \quad (4.165)$$

over the simplex

$$\{\mathbf{w} : w_i > 0 (i = 1, \dots, m); \sum_{i=1}^m w_i = 1\}.$$

It can be seen that (4.165) is a member of the regular exponential family.

Let $\mathbf{w}_1, \dots, \mathbf{w}_n$ denote the observed values of a random sample of size n from the Dirichlet distribution (4.165). The problem is to find the MLE of the parameter vector

$$\boldsymbol{\Psi} = (\theta_1, \dots, \theta_m)^T$$

on the basis of the observed data

$$\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T.$$

An obvious choice for the complete-data vector is

$$\mathbf{x} = (\mathbf{u}_1^T, \dots, \mathbf{u}_n^T)^T,$$

where

$$\mathbf{u}_j = (u_{1j}, \dots, u_{mj})^T$$

and where corresponding to (4.165), the u_{ij} are defined by

$$w_{ij} = u_{ij} / \sum_{h=1}^m u_{hj} \quad (i = 1, \dots, m; j = 1, \dots, n),$$

and $w_{ij} = (\mathbf{w}_j)_i$.

The log likelihood function is given by

$$\begin{aligned} \log L(\boldsymbol{\Psi}) &= \sum_{i=1}^m \sum_{j=1}^n (\theta_i - 1) \log w_{ij} \\ &\quad + n \log \Gamma(\sum_{i=1}^m \theta_i) - n \sum_{i=1}^m \log \Gamma(\theta_i), \end{aligned} \quad (4.166)$$

while the complete-data log likelihood is

$$\begin{aligned}\log L_c(\Psi) &= - \sum_{i=1}^m \sum_{j=1}^n \{(\theta_i - 1) \log u_{ij} - u_{ij}\} \\ &\quad - n \sum_{i=1}^m \log \Gamma(\theta_i).\end{aligned}\quad (4.167)$$

On the E-step at the $(k + 1)$ th iteration of the EM algorithm applied to this problem, we have that

$$\begin{aligned}Q(\Psi; \Psi^{(k)}) &= \sum_{i=1}^m \sum_{j=1}^n \{(\theta_i - 1) \sum_{i=1}^m E_{\Psi^{(k)}}(\log U_{ij} | \mathbf{w}_j) - E_{\Psi^{(k)}}(U_{ij} | \mathbf{w}_j)\} \\ &\quad - n \sum_{i=1}^m \log \Gamma(\theta_i).\end{aligned}\quad (4.168)$$

It can be seen from (4.168) that, in order to carry out the M-step, we need to calculate the term

$$E_{\Psi^{(k)}}(\log U_{ij} | \mathbf{w}_j).$$

Now

$$u_{ij} = w_{ij} \sum_{h=1}^m u_{hj},$$

and so

$$E_{\Psi^{(k)}}(\log U_{ij} | \mathbf{w}_j) = \log w_{ij} + E_{\Psi^{(k)}}(\log U_{.j} | \mathbf{w}_j), \quad (4.169)$$

where

$$U_{.j} = \sum_{h=1}^m U_{hj}$$

is distributed independently of \mathbf{w}_j according to a gamma($\sum_{h=1}^m \theta_h$, 1) density. Thus in order to compute the second term on the right-side of (4.169), we need the result that if a random variable R has a gamma(α, β) distribution, then

$$E(\log R) = \psi(\alpha) - \log \beta, \quad (4.170)$$

where

$$\begin{aligned}\psi(s) &= \partial \log \Gamma(s) / \partial s \\ &= \{\partial \Gamma(s) / \partial s\} / \Gamma(s)\end{aligned}$$

is the Digamma function. On using this result, we have from (4.169) that

$$E_{\Psi^{(k)}}(\log U_{ij} | \mathbf{w}_j) = \log w_{ij} + \psi\left(\sum_{h=1}^m \theta_h^{(k)}\right). \quad (4.171)$$

Lange (1995b) notes that the calculation of the term (4.169) can be avoided if one makes use of the identity

$$S(\mathbf{y}; \Psi^{(k)}) = [\partial Q(\Psi; \Psi^{(k)}) / \partial \Psi]_{\Psi=\Psi^{(k)}}.$$

On evaluating $S_i(\mathbf{y}; \Psi^{(k)})$, the derivative of (4.166) with respect to θ_i at the point $\Psi = \Psi^{(k)}$, we have that

$$\begin{aligned} S_i(\mathbf{y}; \Psi^{(k)}) &= \partial \log L(\Psi^{(k)}) / \partial \theta_i \\ &= \sum_{j=1}^n \log w_{ij} + n \partial \log \Gamma(\sum_{h=1}^m \theta_h^{(k)}) / \partial \theta_i - n \partial \log \Gamma(\theta_i^{(k)}) / \partial \theta_i \\ &= \sum_{j=1}^n \log w_{ij} + n \psi(\sum_{h=1}^m \theta_h^{(k)}) - n \psi(\theta_i^{(k)}) \end{aligned} \quad (4.172)$$

for $i = 1, \dots, m$. On equating $S_i(\mathbf{y}; \Psi^{(k)})$ equal to the derivative of (4.168) with respect to θ_i at the point $\Psi = \Psi^{(k)}$, we obtain

$$\begin{aligned} \sum_{j=1}^n E_{\Psi^{(k)}}(\log U_{ij} | \mathbf{w}_j) &= S_i(\mathbf{y}; \Psi^{(k)}) + n \psi(\theta_i^{(k)}) \\ &= \sum_{j=1}^n \log w_{ij} + n \psi(\sum_{h=1}^m \theta_h^{(k)}), \end{aligned} \quad (4.173)$$

which agrees with sum of the right-hand side of (4.171) as obtained above by working directly with the conditional distribution of $\log U_{ij}$ given \mathbf{w}_j .

Concerning now the M-step, it can be seen that the presence of terms like $\log \Gamma(\theta_i)$ in (4.168) prevent a closed form solution for $\Psi^{(k+1)}$. Lange (1995b) notes that the EM gradient algorithm is easy to implement for this problem. The components of the score statistic $S(\mathbf{y}; \Psi^{(k)})$ are available from (4.172). The information matrix $\mathcal{I}_c(\Psi^{(k)}; \mathbf{y})$, which equals $\mathcal{I}_c(\Psi^{(k)})$, since the complete-data density belongs to the regular exponential family with natural parameter Ψ , is diagonal with i th diagonal element

$$-n \partial^2 \log \Gamma(\theta_i) / \partial \theta_i^2 \quad (i = 1, \dots, m),$$

which is positive as $\log \Gamma(u)$ is a strictly concave function.

Lange (1995b) applied the EM gradient and the quasi-Newton accelerated EM algorithms to a data set from Mosimann (1962) on the relative frequencies of $m = 3$ serum proteins in $n = 23$ young white Pekin ducklings. Starting from $\Psi^{(0)} = (1, 1, 1)^T$, Lange (1995b) reports that the three algorithms converge smoothly to the point $\hat{\Psi} = (3.22, 20.38, 21.69)^T$, with the log likelihood showing a steady increase along the way. The EM gradient algorithm took 287 iterations, while the quasi-Newton accelerated EM algorithm took 8 iterations. Lange (1995b) notes that the scoring algorithm is an attractive alternative for this problem (Narayanan, 1991), and on application to this data set, it took nine iterations.

Table 4.9 gives some details on the performance of the quasi-Newton acceleration of the EM algorithm for this problem. The column headed “Exponent” refers to the minimum nonnegative integer m required to make

$$\mathcal{I}_c(\Psi^{(k)}; \mathbf{y}) + (\frac{1}{2})^m \mathbf{B}^{(k)}, \quad (4.174)$$

positive definite. The column headed “Extra” refers to the number of step decrements taken in order to produce an increase in $L(\Psi)$ at a given iteration. It can be seen that in this problem, step decrementing was not necessary.

Table 4.9 Performance of Accelerated EM on the Dirichlet Distribution Data of Mosimann (1962).

Iteration	Extra	Exponent	$L(\boldsymbol{\theta})$	θ_1	θ_2	θ_3
1	0	0	15.9424	1.000	1.000	1.000
2	0	0	24.7300	0.2113	1.418	1.457
3	0	0	41.3402	0.3897	2.650	2.760
4	0	1	49.1425	0.6143	3.271	3.445
5	0	1	53.3627	0.8222	3.827	4.045
6	0	0	73.0122	3.368	22.19	23.59
7	0	2	73.0524	3.445	22.05	23.47
8	0	0	73.1250	3.217	20.40	21.70
9	0	0	73.1250	3.217	20.39	21.69
10	0	0	73.1250	3.125	20.38	21.69

Source: Adapted from Lange (1995b).

4.15 IKEDA ACCELERATION

Ikeda (2000) proposes an interesting acceleration method which is a hybrid of EM and Fisher's scoring method. It exploits the faster convergence rate of the scoring method. The Ikeda algorithm consists of two steps: (1) Application of an EM step with the given data; (2) Drawing of a random sample from the model with current parameters and application of another EM step with this drawn data.

Let $\tilde{\Psi}^{(k+1)}$ be obtained from $\Psi^{(k)}$ by application of one EM iteration. Then draw a random sample \mathbf{y}^* from the incomplete-data density $g(\mathbf{y}; \Psi^{(k+1)})$ of the same size as data \mathbf{y} . Use \mathbf{y}^* and $\Psi^{(k+1)}$ in one step of the EM algorithm to produce $\tilde{\Psi}^{(k+1)}$. According to the calculations in Ikeda (2000),

$$\tilde{\Psi}^{(k+1)} - \Psi^{(k)} \approx [\mathcal{I}_c^{-1}(\Psi) \mathcal{I}(\Psi) \mathcal{I}_c^{-1}(\Psi) S(\mathbf{y}; \Psi)]_{\Psi=\Psi^{(k)}} \quad (4.175)$$

and

$$\begin{aligned} \tilde{\Psi}^{(k+i)} - \Psi^{(k)} &\approx [\{\mathcal{I}_c^{-1}(\Psi) \mathcal{I}(\Psi)\}^i \mathcal{I}_c^{-1}(\Psi) S(\mathbf{y}; \Psi)]_{\Psi=\Psi^{(k)}} \\ &= [\{\mathbf{I}_d - \mathcal{I}_c^{-1}(\Psi) E_{\Psi}(\mathcal{I}_m(\Psi; \mathbf{Y}))\}^i \mathcal{I}_c^{-1}(\Psi) S(\mathbf{y}; \Psi)]_{\Psi=\Psi^{(k)}}. \end{aligned} \quad (4.176)$$

Ikeda(2000) demonstrates the application of this algorithm for the log linear model and a mixture of normal distributions.

This Page Intentionally Left Blank

EXTENSIONS OF THE EM ALGORITHM

5.1 INTRODUCTION

In this chapter, we consider some extensions of the EM algorithm. In particular, we focus on the ECM and ECME algorithms. The ECM algorithm as proposed by Meng and Rubin (1993), is a natural extension of the EM algorithm in situations where the maximization process on the M-step is relatively simple when conditional on some function of the parameters under estimation. The ECM algorithm therefore replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps. As a consequence, it typically converges more slowly than the EM algorithm in terms of number of iterations, but can be faster in total computer time. More importantly, the ECM algorithm preserves the appealing convergence properties of the EM algorithm, such as its monotone convergence.

Liu and Rubin (1994) propose the ECME algorithm, which is an extension of the ECM algorithm. They find it to be nearly always faster than both the EM and ECM algorithms in terms of number of iterations and moreover that it can be faster in total computer time by orders of magnitude. This improvement in speed of convergence is obtained by conditionally maximizing on some or all of the CM-steps the actual (that is, the incomplete-data) log likelihood rather than a current approximation to it as given by the Q -function with the EM and ECM algorithms. Thus in general it is more tedious to code than the ECM algorithm, but the potential gain in faster convergence allows convergence to be more easily assessed. As with the EM and ECM algorithms, the ECME algorithm monotonically increases the likelihood and reliably converges to a local maximizer of the likelihood function. Several

illustrative examples of the ECM and ECME algorithms are given in this chapter. A further development to be considered is the AECM algorithm of Meng and van Dyk (1997), which is obtained by combining the ECME algorithm with the Space-Alternating Generalized EM (SAGE) algorithm of Fessler and Hero (1994). It allows the specification of the complete data to vary where necessary over the CM-steps. We consider also in this chapter the other proposal of Meng and van Dyk (1997), which concerns speeding up convergence of the EM algorithm through the choice of the complete data.

Liu et al. (1998) developed the Parameter-Expanded EM algorithm (PX-EM) as a means of speeding up the EM algorithm. In this method, the given model is embedded in an expanded model with an additional parameter, say α , such that when α equals a specified value α_0 , the original model is obtained. The algorithm is the same as EM applied to the expanded model. In multivariate incomplete-data models, the gains in speed are substantial.

5.2 ECM ALGORITHM

5.2.1 Motivation

As noted earlier, one of the major reasons for the popularity of the EM algorithm is that the M-step involves only complete-data ML estimation, which is often computationally simple. But if the complete-data ML estimation is rather complicated, then the EM algorithm is less attractive because the M-step is computationally unattractive. In many cases, however, complete-data ML estimation is relatively simple if maximization is undertaken conditional on some of the parameters (or some functions of the parameters). To this end, Meng and Rubin (1993) introduce a class of generalized EM algorithms, which they call the ECM algorithm for expectation-conditional maximization algorithm. The ECM algorithm takes advantage of the simplicity of complete-data conditional maximization by replacing a complicated M-step of the EM algorithm with several computationally simpler CM-steps. Each of these CM-steps maximizes the conditional expectation of the complete-data log likelihood function found in the preceding E-step subject to constraints on Ψ , where the collection of all constraints is such that the maximization is over the full parameter space of Ψ .

A CM-step might be in closed form or it might itself require iteration, but because the CM maximizations are over smaller dimensional spaces, often they are simpler, faster, and more stable than the corresponding full maximizations called for on the M-step of the EM algorithm, especially when iteration is required.

5.2.2 Formal Definition

To define formally the ECM algorithm, we suppose that the M-step is replaced by $S > 1$ steps. We let $\Psi^{(k+s/S)}$ denote the value of Ψ on the s th CM-step of the $(k+1)$ th iteration, where $\Psi^{(k+s/S)}$ is chosen to maximize

$$Q(\Psi; \Psi^{(k)})$$

subject to the constraint

$$\mathbf{g}_s(\Psi) = \mathbf{g}_s(\Psi^{(k+(s-1)/S)}). \quad (5.1)$$

Here $C = \{\mathbf{g}_s(\Psi), s = 1, \dots, S\}$ is a set of S preselected (vector) functions. Thus $\Psi^{(k+s/S)}$ satisfies

$$Q(\Psi^{(k+s/S)}; \Psi^{(k)}) \geq Q(\Psi; \Psi^{(k)}) \quad \text{for all } \Psi \in \Omega_s(\Psi^{(k+(s-1)/S)}), \quad (5.2)$$

where

$$\Omega_s(\Psi^{(k+(s-1)/S)}) \equiv \{\Psi \in \Omega : g_s(\Psi) = g_s(\Psi^{(k+(s-1)/S)})\}. \quad (5.3)$$

The value of Ψ on the final CM-step, $\Psi^{(k+S/S)} = \Psi^{(k+1)}$, is taken to be the input on the $(k+2)$ th iteration.

From (5.2), we have that

$$\begin{aligned} Q(\Psi^{(k+1)}; \Psi^{(k)}) &\geq Q(\Psi^{(k+(S-1)/S)}; \Psi^{(k)}) \\ &\geq Q(\Psi^{(k+(S-2)/S)}; \Psi^{(k)}) \\ &\quad \vdots \\ &\geq Q(\Psi^{(k)}; \Psi^{(k)}). \end{aligned} \quad (5.4)$$

This shows that the ECM algorithm is a GEM algorithm and so possesses its desirable convergence properties. As noted before in Section 3.1, the inequality (5.4) is a sufficient condition for

$$L(\Psi^{(k+1)}) \geq L(\Psi^{(k)})$$

to hold.

Under the assumption that $g_s(\Psi)$, $s = 1, \dots, S$, is differentiable and that the corresponding gradient $\nabla g_s(\Psi)$ is of full rank at $\Psi^{(k)}$, for all k , almost all of the convergence properties of the EM established in DLR and Wu (1983) hold. The only extra condition needed is the “space-filling” condition:

$$\bigcap_{s=1}^S G_s(\Psi^{(k)}) = \{\mathbf{0}\} \quad \text{for all } k, \quad (5.5)$$

where $G_s(\Psi)$ is the column space of $\nabla g_s(\Psi)$; that is,

$$G_s(\Psi) = \{\nabla g_s(\Psi)\eta : \eta \in I\!\!R^{d_s}\}$$

and d_s is the dimensionality of the vector function $g_s(\Psi)$. By taking the complement of both sides of (5.5), this condition is equivalent to saying that at any $\Psi^{(k)}$, the convex hull of all feasible directions determined by the constraint spaces $\Omega_s(\Psi^{(k+(s-1)/S)})$, $s = 1, \dots, S$, is the whole Euclidean space $I\!\!R^d$, and thus the resulting maximization is over the whole parameter space Ω and not a subspace of it. Note that the EM algorithm is a special case of the ECM algorithm with $S = 1$ and $g_1(\Psi) \equiv \text{constant}$ (that is, no constraint), whereby (5.5) is automatically satisfied because

$$\nabla g_1(\Psi) \equiv \mathbf{0}.$$

In many applications of the ECM algorithm, the S CM-steps correspond to the situation where the parameter vector Ψ is partitioned into S subvectors,

$$\Psi = (\Psi_1^T, \dots, \Psi_S^T)^T.$$

The s th CM-step then requires the maximization of the Q -function with respect to the s th subvector Ψ_s with the other $(S-1)$ subvectors held fixed at their current values; that is, $g_s(\Psi)$ is the vector containing all the subvectors of Ψ except Ψ_s ($s = 1, \dots, S$). This is the situation in two of the three illustrative examples that shall be given shortly. In each of these examples, the M-step is not simple, being iterative, and it is replaced by a number of CM-steps which either exist in closed form or require a lower dimensional search.

5.2.3 Convergence Properties

In the case where the data are complete so that the E-step becomes an identity operation, that is,

$$Q(\Psi; \Psi^{(k)}) \equiv L(\Psi),$$

an ECM algorithm becomes a CM algorithm. Meng and Rubin (1993) mention two related issues. First, if the set of constraint functions C is not space-filling, then the CM algorithm will converge to a stationary point of the likelihood in a subspace of Ω , which may or may not be a stationary point of the likelihood in the whole parameter space. Second, since the space-filling condition on C does not involve data, one would expect that if C leads to appropriate convergence of a CM algorithm with complete data, it should also lead to appropriate convergence of an ECM algorithm with incomplete data. This conjecture is established rigorously by Meng and Rubin (1993) when the complete-data density is from an exponential family, where the ECM algorithm is especially useful. The advantage of this is that it enables one to conclude that the ECM algorithm will converge appropriately, whenever the CM does so. For instance, one can immediately conclude the appropriate convergence of the ECM algorithm in Example 5.3 to follow without having to verify the space-filling condition, because the monotone convergence of Iterative Proportional Fitting with complete data has been established (Bishop, Fienberg, and Holland, 2007, Chapter 3).

Meng and Rubin (1993) show that if all the conditional maximizations of an ECM algorithm are unique, then all the limit points of any ECM sequence $\{\Psi^{(k)}\}$ are stationary points of $L(\Psi)$ if C is space-filling at all $\Psi^{(k)}$. The assumption that all conditional maximizations are unique is very weak in the sense that it is satisfied in many practical problems. But even this condition can be eliminated; see Meng and Rubin (1993, page 275). They also provide the formal work that shows that the ECM algorithm converges to a stationary point under essentially the same conditions that guarantee the convergence of the EM algorithm as considered in Chapter 3. Intuitively, suppose that the ECM algorithm has converged to some limit point Ψ^* and that the required derivatives of the Q -function are all well defined. Then the stationarity of each CM-step implies that the corresponding directional derivatives of Q at Ψ^* are zero, which, under the space-filling condition on the set C of constraints, implies that

$$[\partial Q(\Psi; \Psi^*)/\partial \Psi]_{\Psi=\Psi^*} = 0,$$

just as with the M-step of the EM algorithm. Thus, as with the theory of the EM algorithm, if the ECM algorithm converges to Ψ^* , then Ψ^* must be a stationary point of the likelihood function $L(\Psi)$.

5.2.4 Speed of Convergence

From (3.69) and (3.73), the (matrix) speed of convergence of the EM algorithm in a neighborhood of a limit point Ψ^* is given by

$$S_{\text{EM}} = I_d - J(\Psi^*) \quad (5.6)$$

$$= \mathcal{I}_c^{-1}(\Psi^*; y) I(\Psi^*; y), \quad (5.7)$$

and its global speed of convergence s_{EM} by the smallest eigenvalue of (5.7). Under mild conditions, Meng (1994) shows that

$$\text{speed of ECM} = \text{speed of EM} \times \text{speed of CM}; \quad (5.8)$$

that is,

$$S_{\text{ECM}} = S_{\text{EM}} S_{\text{CM}}, \quad (5.9)$$

where S_{CM} and S_{ECM} denote the speed of the CM and ECM algorithms, respectively, corresponding to (5.7). This result is consistent with intuition, since an ECM iteration can be viewed as a composition of two linear iterations, EM and CM.

Let s_{ECM} and s_{CM} denote the global speed of convergence of the ECM and CM algorithms, respectively. As Meng (1994) comments, although it would be naive to expect from (5.9) that

$$s_{\text{ECM}} = s_{\text{EM}} s_{\text{CM}}, \quad (5.10)$$

it seems intuitive to expect that

$$s_{\text{EM}} s_{\text{CM}} \leq s_{\text{ECM}} \leq s_{\text{EM}}. \quad (5.11)$$

Because the increase in the Q -function is less with the ECM algorithm than with the EM algorithm, it would appear reasonable to expect that the former algorithm converges more slowly in terms of global speed, which would imply the upper bound on s_{ECM} in (5.11). But intuition suggests that the slowest convergence of the ECM algorithm occurs when the EM and CM algorithms share a slowest component. In this case, the speed of the ECM algorithm is the product of the speeds of the EM and CM algorithms, leading to the lower bound on s_{ECM} in (5.11). However, neither of the inequalities in (5.11) holds in general, as Meng (1994) uses a simple bivariate normal example to provide counterexamples to both inequalities.

5.2.5 Convergence Rates of EM and ECM

A penalty to pay for the added simplicity and stability of the ECM algorithm is its slower convergence relative to EM, especially if the ECM is implemented in a natural manner. However, Sexton and Swensen (2000) show that if the CM-steps are implemented to satisfy a certain property then the rate of convergence of ECM can always be made equal or approximately equal to that of EM.

Suppose in the s th CM-step following a k th E-step $Q(\Psi; \Psi^{(k)})$ is maximized over a set of vectors $\mathbf{d}_s^{(1)}, \dots, \mathbf{d}_s^{(m_s)}$, where the $\mathbf{d}_s^{(h)} (h = 1, \dots, m_s)$ are column vectors such that

$$\mathbf{d}_s = (\mathbf{d}_1^{(1)}, \dots, \mathbf{d}_s^{(m_s)})$$

is of rank m_s . They are chosen so that

$$\mathbf{d}_i^{(h_i)^T} \mathcal{I}_c(\Psi^*; \mathbf{y}) \mathbf{d}_j^{(h_j)} = 0 \quad i \neq j; i, j \in \{1, \dots, S\}. \quad (5.12)$$

Such vectors are called \mathcal{I}_c -orthogonal. We use the notation D for the differential operator, like DM for the differential of the M -function. Then, Sexton and Swensen (2000) prove the following results that show the relationship between the rates of convergence of EM and ECM (exactly and approximately respectively).

If the vectors searched over in each CM-step are \mathcal{I}_c -orthogonal to the vectors searched over in the other CM-steps then

$$DM^{\text{ECM}} = DM^{\text{EM}}.$$

If the vectors searched over in each CM-step are asymptotically \mathcal{I}_c -orthogonal, that is, for $j \neq k$,

$$\frac{1}{n} \mathbf{d}_j^{(h_j)^T} \mathcal{I}_c^{(n)} \mathbf{d}_k^{(h_k)} \rightarrow 0$$

in probability as $n \rightarrow \infty$, where n denotes the number of observations and $\mathcal{I}_c^{(n)}$ is the $\mathcal{I}_c(\Psi; \mathbf{y})$ matrix with n observations evaluated at the MLE from these observations, then

$$\| D\mathcal{M}^{\text{ECM}} - D\mathcal{M}^{\text{EM}} \| \rightarrow 0$$

in probability as $n \rightarrow \infty$.

5.2.6 Example 5.1: ECM Algorithm for Hidden Markov AR(1) Model

Continuing with the example of a hidden Markov AR(1) model (2.81) discussed in Section 2.9, suppose for the sake of illustration we assume that the transition probabilities π_{hi} of the Markov chain are known to be all equal to 0.5, that $\mu_1 = 0$, $\sigma^2 = 1$, and that the autoregressive parameters β_i are the same in each state ($\beta_1 = \beta_2 = \beta$). This leaves $\Psi = (\mu_2, \beta)^T$ as the vector of unknown parameters. Also, it is assumed that $\text{pr}(W_1 = 0, s_1 = 1) = 1$.

The ECM algorithm is useful in this case, since if μ_2 is known then the estimation of the autoregressive parameter β can be performed analytically. These parameters are asymptotically \mathcal{I}_c -orthogonal and so the ECM algorithm converges at approximately the same rate as the EM algorithm and is easier to implement.

The ECM algorithm is as follows:

E-Step. The step requires the derivation of the $Q(\Psi; \Psi^k)$ function, which needs the computation of the probabilities $\tau_{hij;n}^{(k)}$ for $j = 2, \dots, n; h, i = 1, 2$, where

$$\tau_{hij;n}^{(k)} = \text{pr}_{\Psi^{(k)}}\{s_{j-1} = h, s_j = i \mid \mathbf{y}_n\}.$$

This can be effected by the smoothing method discussed in Section 2.9.

CM-Steps.

$$\mu_2^{(k+1)} = \frac{\sum_{j=2}^n (w_j - \beta^{(k)} w_{j-1}) [\tau_{21;j}^{(k)} - \beta^{(k)} \tau_{12;j}^{(k)} + (1 - \beta^{(k)}) \tau_{22;j}^{(k)}]}{\sum_{j=2}^n [(\beta^{(k)})^2 \tau_{12;j}^{(k)} + \tau_{21;j}^{(k)} + (1 - \beta^{(k)}) \tau_{22;j}^{(k)}]} \quad (5.13)$$

$$\beta^{(k+1)} = \frac{\sum_{j=2}^n \sum_{h,i=1}^2 (w_{j-1} - \mu_{s_{j-1}}^{(k+1)}) (w_j - \mu_{s_j}^{(k+1)}) \tau_{hi;j}^{(k)}}{\sum_{j=2}^n \sum_{i=1}^2 (y_{j-1} - \mu_{s_{j-1}}^{(k+1)})^2 \tau_{ij}^{(k)}}, \quad (5.14)$$

where $\tau_{ij}^{(k)} = \text{pr}_{\Psi^{(k)}}\{s_j = i \mid \mathbf{y}_n\}$; see Sexton and Swensen (2000) for more details.

5.2.7 Discussion

The CM algorithm is a special case of the cyclic coordinate ascent method for function maximization in the optimization literature; see, for example, Zangwill (1969, Chapter 5). It can also be viewed as the Gauss-Seidel iteration method applied in an appropriate order to the likelihood equation (for example, Thisted, 1988, Chapter 4). Although these optimization methods are well known for their simplicity and stability, because they typically converge only linearly, they have been less preferred in practice for handling complete-data problems than superlinear methods like quasi-Newton. When used for the M-step of the EM algorithm or a CM-step of the ECM algorithm, however, simple and stable linear converging methods are often more suitable than superlinear converging but less stable algorithms. The

reasons are first, that the advantage of superlinear convergence in each M- or CM-step does not transfer to the overall convergence of the EM or ECM algorithms, since the EM and ECM algorithms always converge linearly regardless of the maximization method employed within the maximization step, and secondly, that the stability of the maximization method is critical for preserving the stability of the EM or ECM algorithms since it is used repeatedly within each maximization step in all iterations. Finally, if one performs just one iteration of a superlinear converging algorithm within each M-step of the EM algorithm, then the resulting algorithm is no longer guaranteed to increase the likelihood monotonically.

As emphasized by Meng (1994), although the ECM algorithm may converge slower than the EM algorithm in terms of global speed of convergence, it does not necessarily imply that it takes longer to converge in real time since the actual time in running the M-step and the CM-steps is not taken into account in the speed of convergence. This time can be quite different, especially if the M-step requires iteration.

5.3 MULTICYCLE ECM ALGORITHM

In many cases, the computation of an E-step may be much cheaper than the computation of the CM-steps. Hence one might wish to perform one E-step before each CM-step or a few selected CM-steps. For descriptive simplicity, we focus here on the case with an E-step preceding each CM-step. A cycle is defined to be one E-step followed by one CM-step. Meng and Rubin (1993) called the corresponding algorithm a multicycle ECM.

For instance, consider an ECM algorithm with $S = 2$ CM-steps. Then an E-step would be performed between the two CM-steps. That is, after the first CM-step, an E-step is performed, by which the Q -function is updated from

$$Q(\Psi; \Psi^{(k)}) \quad (5.15)$$

to

$$Q(\Psi; \Psi^{(k+1/2)}). \quad (5.16)$$

The second CM-step of this multicycle ECM algorithm is then undertaken where now $\Psi^{(k+2/S)} = \Psi^{(k+2)}$ is chosen to maximize conditionally (5.16) instead of (5.15).

Since the second argument in the Q -function is changing at each cycle within each iteration, a multicycle ECM may not necessarily be a GEM algorithm; that is, the inequality (5.4) may not be hold. However, it is not difficult to show that the likelihood function is not decreased after a multicycle ECM iteration, and hence, after each ECM iteration. To see this, we have that the definition of the s th CM-step implies

$$Q(\Psi^{(k+s/S)}; \Psi^{(k+(s-1)/S)}) \geq Q(\Psi^{(k+(s-1)/S)}; \Psi^{(k+(s-1)/S)}),$$

which we have seen is sufficient to establish that

$$L(\Psi^{(k+s/S)}) \geq L(\Psi^{(k+(s-1)/S)}).$$

Hence the multicycle ECM algorithm monotonically increases the likelihood function $L(\Psi)$ after each cycle, and hence, after each iteration. The convergence results of the ECM algorithm apply to a multicycle version of it. An obvious disadvantage of using a multicycle ECM algorithm is the extra computation at each iteration. Intuitively, as a tradeoff, one might expect it to result in larger increases in the log likelihood function per iteration since the Q -function is being updated more often. Meng and Rubin (1993) report that practical

implementations do show this potential, but it is not true in general. They note that there are cases where the multicycle ECM algorithm converges more slowly than the ECM algorithm. Details of these examples, which are not typical in practice, are given in Meng (1994).

We shall now give some examples, illustrating the application of the ECM algorithm and its multicycle version. The first few examples show that when the parameters are restricted to particular subspaces, the conditional maximizations either have analytical solutions or require lower dimensional iteration.

5.4 EXAMPLE 5.2: NORMAL MIXTURES WITH EQUAL CORRELATIONS

5.4.1 Normal Components with Equal Correlations

As an example of the application of the ECM algorithm, we consider the fitting of a mixture of two multivariate normal densities with equal correlation matrices to an observed random sample $\mathbf{w}_1, \dots, \mathbf{w}_n$. In Section 2.7.2 we considered the application of the EM algorithm to fit a mixture of g normal component densities with unrestricted covariance matrices.

The case of two equal component-correlation matrices can be represented by writing the covariance matrices Σ_1 and Σ_2 as

$$\Sigma_1 = \Sigma_0 \quad (5.17)$$

and

$$\Sigma_2 = \mathbf{K} \Sigma_0 \mathbf{K}, \quad (5.18)$$

where Σ_0 is a positive definite symmetric matrix and

$$\mathbf{K} = \text{diag}(\kappa_1, \dots, \kappa_p).$$

For the application of the ECM algorithm, the vector Ψ of unknown parameters can be partitioned as

$$\Psi = (\Psi_1^T, \Psi_2^T)^T, \quad (5.19)$$

where Ψ_1 , consists of the mixing proportion π_1 , the elements of μ_1 and μ_2 , and the distinct elements of Σ_0 , and where Ψ_2 contains $\kappa_1, \dots, \kappa_p$.

5.4.2 Application of ECM Algorithm

The E-step on the $(k + 1)$ th iteration of this ECM algorithm is the same as for the EM algorithm. In this particular case where the two CM-steps correspond to the maximization of the Q -function with respect to each of the two subvectors in the partition (5.19) of Ψ , we can write the values of Ψ obtained after each CM-step as

$$\Psi^{(k+1/2)} = (\Psi_1^{(k+1)^T}, \Psi_2^{(k)^T})^T$$

and

$$\Psi^{(k+1)} = (\Psi_1^{(k+1)^T}, \Psi_2^{(k+1)^T})^T,$$

respectively. The two CM-steps of the ECM algorithm can then be expressed as follows.

CM-Step 1. Calculate $\Psi_1^{(k+1)}$ as the value of Ψ_1 that maximizes $Q(\Psi; \Psi^{(k)})$ with Ψ_2 fixed at

$$\Psi_2^{(k)} = (\kappa_1^{(k)}, \dots, \kappa_p^{(k)})^T.$$

CM-Step 2 Calculate $\Psi_2^{(k+1)}$ as the value of Ψ_2 that maximizes $Q(\Psi; \Psi^{(k)})$ with Ψ_1 fixed at $\Psi_1^{(k+1)}$.

The CM-Step 1 can be implemented by proceeding as with the M-step in Section 2.7. There is a slight modification to allow for the fact that the two component-covariance matrices are related by (5.17) and (5.18); that is, have the same correlation structure. It follows that $\Psi_1^{(k+1)}$ consists of $\pi_1^{(k+1)}$, the elements of $\mu_1^{(k+1)}$ and $\mu_2^{(k+1)}$, and the distinct elements of $\Sigma_0^{(k+1)}$, where

$$\pi_i^{(k+1)} = \sum_{j=1}^n z_{ij}^{(k)} / n \quad (i = 1, 2),$$

$$\mu_i^{(k+1)} = \sum_{j=1}^n z_{ij}^{(k)} \mathbf{w}_j / \sum_{j=1}^n z_{ij}^{(k)} \quad (i = 1, 2),$$

and

$$\Sigma_0^{(k+1)} = \pi_1^{(k+1)} \mathbf{V}_1^{(k+1)} + \pi_2^{(k+1)} \mathbf{K}^{(k)^{-1}} \mathbf{V}_2^{(k+1)} \mathbf{K}^{(k)^{-1}},$$

where

$$\mathbf{V}_i^{(k+1)} = \sum_{j=1}^n z_{ij}^{(k)} (\mathbf{w}_j - \mu_i^{(k+1)}) (\mathbf{w}_j - \mu_i^{(k+1)})^T / \sum_{j=1}^n z_{ij}^{(k)} \quad (i = 1, 2), \quad (5.20)$$

and

$$\mathbf{K}^{(k)} = \text{diag}(\kappa_1^{(k)}, \dots, \kappa_p^{(k)})^T.$$

As in Section 2.7,

$$z_{ij}^{(k)} = \tau_i(\mathbf{w}_j; \Psi^{(k)}) \quad (i = 1, 2),$$

but here now

$$\tau_1(\mathbf{w}_j; \Psi^{(k)}) = \frac{\pi_1^{(k)} \phi(\mathbf{w}_j; \mu_1^{(k)}, \Sigma_0^{(k)})}{\pi_1^{(k)} \phi(\mathbf{w}_j; \mu_1^{(k)}, \Sigma_0^{(k)}) + \pi_2^{(k)} \phi(\mathbf{w}_j; \mu_2^{(k)}, \Sigma_2^{(k)})},$$

where

$$\Sigma_2^{(k)} = \mathbf{K}^{(k)} \Sigma_0^{(k)} \mathbf{K}^{(k)},$$

and

$$\tau_2(\mathbf{w}_j; \Psi^{(k)}) = 1 - \tau_1(\mathbf{w}_j; \Psi^{(k)}).$$

Concerning CM-Step 2, it is not difficult to show (McLachlan, 1992, Chapter 5)) that $\kappa_v^{(k+1)}$ is a solution of the equation

$$\kappa_v = \sum_{i=1}^p \left(\Sigma_0^{(k+1)^{-1}} \right)_{iv} \left(\mathbf{V}_2^{(k+1)} \right)_{iv} / \kappa_i \quad (v = 1, \dots, p). \quad (5.21)$$

This equation can be solved iteratively (for example, by Newton-Raphson or a quasi-Newton procedure) to yield $\kappa_v^{(k+1)}$ ($v = 1, \dots, p$) and hence $\mathbf{K}^{(k+1)}$; that is, $\Psi_2^{(k+1)}$.

We may wish to consider a multicycle version of this ECM algorithm, where an E-step is introduced between the two CM-steps. The effect of this additional E-step is to replace

$$z_{ij}^{(k)} = \tau_i(\mathbf{w}_j; \boldsymbol{\Psi}^{(k)})$$

by $\tau_i(\mathbf{w}_j; \boldsymbol{\Psi}^{(k+1/S)})$ in forming $\mathbf{V}_2^{(k+1)}$ from (5.20) for use in (5.21).

5.4.3 Fisher's Iris Data

McLachlan (1992, Chapter 6) and McLachlan , Basford, and Green (1993) fitted a mixture of two normals with unrestricted covariance matrices to this data set in order to assess whether the *virginica* species in Fisher's well-known and analyzed *Iris* data should be split into two subspecies. One of their solutions corresponding to a local maximum of the likelihood function produces a cluster containing five observations numbered 6, 18, 19, 23, and 31, with the remaining 45 observations in a second cluster. The observations are labeled 1 to 50 in order of their listing in Table 1.1 in Andrews and Herzberg (1985). In biological applications involving clusters, it is often reasonable to assume that the correlation matrices are the same within each cluster. Therefore, in order to avoid potential problems with spurious local maxima as a consequence of clusters having a very small generalized variance, McLachlan and Prado (1995) imposed the restriction of equal correlation matrices of the normal components of the mixture model fitted to this data set. The effect of this restriction on the two clusters above for unrestricted component-covariance matrices was to move the entity numbered 31 to the larger cluster.

5.5 EXAMPLE 5.3: MIXTURE MODELS FOR SURVIVAL DATA

5.5.1 Competing Risks in Survival Analysis

In this example, we illustrate the use of the ECM algorithm for ML fitting of finite mixture models in survival analysis. We suppose that the p.d.f. of the failure time T for a patient has the finite mixture representation

$$f(t; a) = \sum_{i=1}^g f_i(t; a), \quad (5.22)$$

where $f_1(t; a), \dots, f_g(t; a)$ denote g component densities occurring in proportions π_1, \dots, π_g , and $0 \leq \pi_i \leq 1$ ($i = 1, \dots, g$). Here a denotes a covariate associated with the patient, for instance, his or her age at some designated time, for example, at the time of therapy for a medical condition.

One situation where the mixture model is directly applicable for the p.d.f. $f(t; a)$ of the failure time T is where the adoption of one parametric family for the distribution of failure time is inadequate. A way of handling this is to adopt a mixture of parametric families. Another situation where the mixture model (5.22) is directly applicable for the p.d.f. of the failure time is where a patient is exposed to g competing risks or types of failure. The p.d.f. of the failure time T has the mixture form (5.22), where $f_i(t; a)$ is the p.d.f. of T conditional on the failure being of type i , and π_i is the prior probability of a type i failure ($i = 1, \dots, g$). In effect, this mixture approach assumes that a patient will fail from a particular risk, chosen by a stochastic mechanism at the outset. For example, after therapy

for lung cancer, an uncured patient might be destined to die from lung cancer, while a cured patient will eventually die from other causes. Farewell (1982, 1986) and Larson and Dinse (1985) were among the first to use finite mixture models to handle competing risks; see McLachlan and McGiffen (1994) for a survey of the use of mixture models in survival analysis.

5.5.2 A Two-Component Mixture Regression Model

In the following, we consider the case of $g = 2$ components in the mixture model (5.22), corresponding to two mutually exclusive groups of patients G_1 and G_2 , where G_1 corresponds to those patients who die from a particular disease and G_2 to those who die from other causes.

The effect of the covariate a (say, age) on the mixing proportions is modeled by the logistic model, under which

$$\pi_1(a; \beta) = e^{\beta_0 + \beta_1 a} / (1 + e^{\beta_0 + \beta_1 a}), \quad (5.23)$$

where $\beta = (\beta_0, \beta_1)^T$ is the parameter vector.

The survivor function,

$$\bar{F}(t; a) = \text{pr}\{T > t | a\},$$

for a patient aged a , can therefore be represented by the two-component mixture model

$$\bar{F}(t; a) = \pi_1(a; \beta)\bar{F}_1(t; a) + \pi_2(a; \beta)\bar{F}_2(t; a), \quad (5.24)$$

where $\bar{F}_i(t; a)$ is the probability that T is greater than t , given that the patient belongs to G_i ($i = 1, 2$).

5.5.3 Observed Data

For patient j ($j = 1, \dots, n$), we observe

$$\mathbf{y}_j = (t_j, a_j, \delta_{1j}, \delta_{2j}, \delta_{3j})^T,$$

where δ_{1j} , δ_{2j} , and δ_{3j} are zero-one indicator variables with $\delta_{1j} = 1$ if patient j died at time t_j (after therapy) for the disease under study and zero otherwise, $\delta_{2j} = 1$ if patient j died from other causes at time t_j and zero otherwise, and $\delta_{3j} = 1$ if patient j was still alive at the termination of the study at time t_j and zero otherwise. Thus

$$\delta_{1j} + \delta_{2j} + \delta_{3j} = 1$$

for $j = 1, \dots, n$. Also, a_j denotes the age of the j th patient. The observed times t_1, \dots, t_n are regarded as n independent observations on the random variable T defined to be the time to death either from the disease under study or from other causes. For the patients who are still living at the end of the study ($\delta_{3j} = 1$), their failure times are therefore censored, where it is assumed that the censoring mechanism is noninformative. That is, the censored observation provides only a lower bound on the failure time, and given that bound, the act of censoring imparts no information about the eventual cause of death.

The two-component mixture model (5.24) can be fitted by ML methods if we specify the component survivor functions $\bar{F}_1(t; a)$ and $\bar{F}_2(t; a)$ up to a manageable number of unknown parameters. We henceforth write $\bar{F}_1(t; a)$ and $\bar{F}_2(t; a)$ as $\bar{F}_1(t; a, \theta_1)$ and $\bar{F}_2(t; a, \theta_2)$,

respectively, where $\boldsymbol{\theta}_i$ denotes the vector of unknown parameters in the specification of the i th component survivor function ($i = 1, 2$). We let

$$\boldsymbol{\Psi} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$$

be the vector containing all the unknown parameters. Then the log likelihood for $\boldsymbol{\Psi}$ formed on the basis of $\mathbf{y}_1, \dots, \mathbf{y}_n$ is given by

$$\begin{aligned} \log L(\boldsymbol{\Psi}) &= \sum_{j=1}^n \left[\sum_{i=1}^2 \delta_{ij} \{ \log \pi_i(a_j; \boldsymbol{\beta}) + \log f_i(t_j; a_j, \boldsymbol{\theta}_i) \} \right. \\ &\quad \left. + \delta_{3j} \log \{ \pi_1(a_j; \boldsymbol{\beta}) \bar{F}_1(t_j; a_j, \boldsymbol{\theta}_1) + \pi_2(a_j; \boldsymbol{\beta}) \bar{F}_2(t_j; a_j, \boldsymbol{\theta}_2) \} \right], \end{aligned} \quad (5.25)$$

where

$$f_i(t; a_j, \boldsymbol{\theta}_i) = -d\bar{F}_i(t; a_j, \boldsymbol{\theta}_i)/dt$$

is the density function of T in G_i ($i = 1, 2$).

5.5.4 Application of EM Algorithm

Rather than working directly with the log likelihood (5.25), the MLE of $\boldsymbol{\Psi}$ can be found by an application of the EM algorithm. If $\delta_{3j} = 1$ for patient j (that is, if the survival time t_j is censored), it is computationally convenient to introduce the zero-one indicator variable z_{ij} , where $\mathbf{z}_j = (z_{1j}, z_{2j})^T$ and $z_{ij} = 1$ or 0 according to whether this patient belongs to group G_i or not ($i = 1, 2$; $j = 1, \dots, n$). We can then apply the EM algorithm within the framework where $(\mathbf{y}_1^T, \mathbf{z}_1^T)^T, \dots, (\mathbf{y}_n^T, \mathbf{z}_n^T)^T$ are viewed as the complete data. The actual time to failure for those patients with $\delta_{3j} = 1$ is not introduced as an incomplete variable in the complete-data framework, as it does not simplify the calculations.

The complete-data log likelihood is given then by

$$\begin{aligned} \log L_c(\boldsymbol{\Psi}) &= \sum_{j=1}^n \left[\sum_{i=1}^2 \delta_{ij} \{ \log \pi_i(a_j; \boldsymbol{\beta}) + \log f_i(t_j; a_j, \boldsymbol{\theta}_i) \} \right. \\ &\quad \left. + \delta_{3j} \sum_{i=1}^2 z_{ij} \{ \log \pi_i(a_j; \boldsymbol{\beta}) + \log \bar{F}_i(t_j; a_j, \boldsymbol{\theta}_i) \} \right]. \end{aligned} \quad (5.26)$$

On the E-step at the $(k+1)$ iteration of the EM algorithm, we have to find the conditional expectation of $\log L_c(\boldsymbol{\Psi})$, given the observed data $\mathbf{y}_1, \dots, \mathbf{y}_n$, using the current fit $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. This is effected by replacing z_{ij} in $\log L_c(\boldsymbol{\Psi})$ by $z_{ij}^{(k)}$, which is its current conditional expectation given \mathbf{y}_j , for each patient j that has $\delta_{3j} = 1$. Now

$$\begin{aligned} z_{ij}^{(k)} &= E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij} | t_j, \delta_{3j} = 1, a_j) \\ &= \tau_i(t_j; a_j, \boldsymbol{\Psi}^{(k)}), \end{aligned} \quad (5.27)$$

where

$$\tau_1(t_j; a_j, \boldsymbol{\Psi}^{(k)}) = \frac{\pi_1(a_j; \boldsymbol{\beta}^{(k)}) \bar{F}_1(t_j; a_j, \boldsymbol{\theta}_1^{(k)})}{\pi_1(a_j; \boldsymbol{\beta}^{(k)}) \bar{F}_1(t_j; a_j, \boldsymbol{\theta}_1^{(k)}) + \pi_2(a_j; \boldsymbol{\beta}^{(k)}) \bar{F}_2(t_j; a_j, \boldsymbol{\theta}_2^{(k)})}$$

and $\tau_2(t_j; a_j, \Psi^{(k)}) = 1 - \tau_1(t_j; a_j, \Psi^{(k)})$.

From above, $Q(\Psi; \Psi^{(k)})$ is given by the expression for $\log L_c(\Psi)$ with z_{ij} replaced by $z_{ij}^{(k)}$ for those patients j with $\delta_{3j} = 1$. The M-step then involves choosing $\Psi^{(k+1)}$ so as to maximize $Q(\Psi; \Psi^{(k)})$ with respect to Ψ to give $\Psi^{(k+1)}$.

5.5.5 M-Step for Gompertz Components

We now pursue the implementation of the M-step in the case where the component survivor functions are modeled by the Gompertz distribution additively adjusted on the log scale for the age a of the patient. That is, $\bar{F}_i(t; a, \theta_i)$ is modeled as

$$\bar{F}_i(t; a, \theta_i) = \exp\{-e^{\lambda_i + \gamma_i a}(e^{\xi_i t} - 1)/\xi_i\}, \quad (5.28)$$

where $\theta_i = (\lambda_i, \xi_i, \gamma_i)^T$ ($i = 1, 2$).

The identifiability of mixtures of Gompertz distributions has been established by Gordon (1990a, 1990b) in the case of mixing proportions that do not depend on any covariates. The extension to the case of mixing proportions specified by the logistic model (5.23) is straightforward. It follows that a sufficient condition for identifiability of the Gompertz mixture model (5.28) is that the matrix $(\mathbf{a}_1^+, \dots, \mathbf{a}_n^+)$ be of full rank, where

$$\mathbf{a}_j^+ = (1, a_j)^T.$$

Unfortunately, for Gompertz component distributions, $\Psi^{(k+1)}$ does not exist in closed form. We consider its iterative computation, commencing with $\beta^{(k+1)}$. On equating the derivative of $Q(\Psi; \Psi^{(k)})$ with respect to β to zero, we have that $\beta^{(k+1)}$ can be computed iteratively by the Newton-Raphson method as

$$\beta^{(k+1, m+1)} = \beta^{(k+1, m)} + \mathbf{B}^{(k+1, m)-1} \mathbf{U}^{(k, m)},$$

where

$$\mathbf{B}^{(k+1, m)} = \mathbf{A}^T \mathbf{V}^{(k+1, m)} \mathbf{A}, \quad (5.29)$$

$$\mathbf{V}^{(k+1, m)} = \text{diag}(v_1^{(k+1, m)}, \dots, v_n^{(k+1, m)}), \quad (5.30)$$

$$v_j^{(k+1, m)} = \pi_1(a_j; \beta^{(k+1, m)}) \pi_2(a_j; \beta^{(k+1, m)}), \quad (5.31)$$

$$\mathbf{A} = (\mathbf{a}_1^+, \dots, \mathbf{a}_n^+)^T,$$

$$\mathbf{U}^{(k, m)} = \mathbf{w}^{(k)} - \sum_{j=1}^n \pi_1(a_j; \beta^{(k+1, m)}) \mathbf{a}_j^+, \quad (5.32)$$

and

$$\mathbf{w}^{(k)} = \sum_{j=1}^n \{\delta_{1j} + \delta_{3j} \tau_1(t_j; a_j, \Psi^{(k)})\} \mathbf{a}_j^+. \quad (5.33)$$

Provided the sequence $\{\beta^{(k+1, m+1)}\}$ converges, as $m \rightarrow \infty$, $\beta^{(k+1)}$ can be taken equal to $\beta^{(k+1, m+1)}$ for m sufficiently large.

On equating the derivatives of $Q(\Psi; \Psi^{(k)})$ with respect to the elements of θ_i , we have that $\lambda_i^{(k+1)}, \xi_i^{(k+1)}$, and $\gamma_i^{(k+1)}$ satisfy the following equations

$$\sum_{j=1}^n \{\delta_{ij} - (\delta_{ij} + \delta_{3j} z_{ij}^{(k)}) (h_{ij}^{(k+1)} / \xi_i^{(k+1)})\} = 0, \quad (5.34)$$

$$\sum_{j=1}^n \{\delta_{ij}t_j + (\delta_{ij} + \delta_{3j}z_{ij}^{(k)})(-t_j c_{ij}^{(k+1)}/\xi_i^{(k+1)} + h_{ij}^{(k+1)}/\xi_i^{(k+1)^2})\} = 0, \quad (5.35)$$

and

$$\sum_{j=1}^n \{\delta_{ij}a_j - (\delta_{ij} + \delta_{3j}z_{ij}^{(k)})a_j h_{ij}^{(k+1)}/\xi_i^{(k+1)}\} = 0 \quad (5.36)$$

for $i = 1$ and 2, where

$$h_{ij}^{(k+1)} = \exp(\lambda_i^{(k+1)} + \gamma_i^{(k+1)}a_j)\{\exp(\xi_i^{(k+1)}t_j) - 1\}$$

and

$$c_{ij}^{(k+1)} = \exp(\lambda_i^{(k+1)} + \gamma_i^{(k+1)}a_j + \xi_i^{(k+1)}t_j).$$

From (5.34), $\lambda_i^{(k+1)}$ can be expressed in terms of $\gamma_i^{(k+1)}$ and $\xi_i^{(k+1)}$ as

$$\lambda_i^{(k+1)} = \log\{q_{1i}^{(k+1)}/q_{2i}^{(k+1)}\} \quad (5.37)$$

for $i = 1$ and 2, where

$$q_{1i}^{(k+1)} = \xi_i^{(k+1)} \sum_{j=1}^n \delta_{ij}$$

and

$$q_{2i}^{(k+1)} = \sum_{j=1}^n (\delta_{ij} + \delta_{3j}z_{ij}^{(k)}) \exp(\gamma_i^{(k+1)}a_j)\{\exp(\xi_i^{(k+1)}t_j) - 1\}.$$

Thus in order to obtain $\theta_i^{(k+1)}$, the two equations (5.35) and (5.36), where $\lambda_i^{(k+1)}$ is given by (5.37), have to be solved. This can be undertaken iteratively using a quasi-Newton method, as in the FORTRAN program of McLachlan, Ng, Adams, McGiffin, and Galbraith (1994).

5.5.6 Application of a Multicycle ECM Algorithm

We partition Ψ as

$$\Psi = (\Psi_1^T, \Psi_2^T)^T,$$

where $\Psi_1 = \beta$ and $\Psi_2 = (\theta_1^T, \theta_2^T)^T$. From above, it can be seen that $\Psi_1^{(k+1)}$ and $\Psi_2^{(k+1)}$ are computed independently of each other on the M-step of the EM algorithm. Therefore, the latter is the same as the ECM algorithm with two CM-steps, where on the first CM-step, $\Psi_1^{(k+1)}$ is calculated with Ψ_2 fixed at $\Psi_2^{(k)}$, and where on the second CM-step, $\Psi_2^{(k+1)}$ is calculated with Ψ_1 fixed at $\Psi_1^{(k+1)}$.

In order to improve convergence, McLachlan et al. (1994) use a multicycle version of this ECM algorithm where an E-step is performed after the computation of $\beta^{(k+1)}$ and before the computation of the other subvectors $\theta_1^{(k+1)}$ and $\theta_2^{(k+1)}$ in $\Psi_2^{(k+1)}$. This multicycle E-step is effected here by updating $\beta^{(k)}$ with $\beta^{(k+1)}$ in $\Psi^{(k)}$ in the right-hand side of the expression (5.27) for $z_{ij}^{(k)}$.

To assist further with the convergence of the sequence of iterates $\beta^{(k+1, m+1)}$, we can also perform an additional E-step after the computation of $\beta^{(k+1, m+1)}$ before proceeding

with the computation of $\beta^{(k+1, m+2)}$ during the iterative computations on the first CM-step. That is, on the right-hand sides of (5.29) to (5.33), $z_{ij}^{(k)} = \tau_1(t_j; a_j, \Psi^{(k)})$ is replaced by

$$\tau_1(t_j; a_j, \Psi^{(k+1, m+1)}),$$

where

$$\Psi^{(k+1, m+1)} = (\beta^{(k+1, m+1)^T}, \theta_1^{(k)^T}, \theta_2^{(k)^T})^T.$$

This is no longer a multicycle ECM algorithm, and so the inequality

$$L(\Psi^{(k+1)}) \geq L(\Psi^{(k)}) \quad (5.38)$$

does not necessarily hold. However, it has been observed to hold for the data sets analyzed by McLachlan, Adams, Ng, McGiffen, and Galbraith (1994).

5.5.7 Other Examples of EM Algorithm in Survival Analysis

There are many more examples of the application of EM algorithm in Survival Analysis. We mention a few of them here.

DeGruttola and Tu (1994) apply the EM algorithm to estimate parameters in a joint model of survival times and disease progression under normally distributed random effects. They assume that conditional on the random effects (of individuals) the outcomes of survival times and disease progression are independent, making it easy to specify the joint likelihood. Standard errors are obtained using Louis' method. Wulfsohn and Tsiatis (1997) use a proportional hazards model for survival times conditional on the longitudinal marker, and model the covariate and measurement error by a random effects model. They also apply the EM algorithm for parameter estimation, using numerical integration in the E-step and Newton-Raphson approximations in the M-step. These examples are described in Ibrahim, Chen and Sinha (2001) who go on to describe a Bayesian approach to joint modeling of survival and longitudinal data.

Klein (1992) and Nielsen, Gill, Andersen, and Sørensen (1992) consider a gamma frailty model due to Clayton (1978) and Clayton and Cuzick (1985) with parameter θ and a survival function with regression parameters β on the covariates. In case no parametric form is formulated for the hazard rate, they obtain semiparametric estimates using the EM algorithm. In the M-step, a partial likelihood is constructed using a suitable Cox model and maximized to update β , whereas updates of θ are obtained by maximizing the full likelihood; See also Gill (1985). Nielsen et al. (1992) develop a counting process derivation of this approach and suggest a modified EM algorithm which converges faster than the above EM by maximizing a profile likelihood. These procedures are discussed in Klein and Moeschberger (1997).

In a series of articles, Balakrishnan and coworkers (Ng, Chan, and Balakrishnan, 2002, 2004; Balakrishnan and Kim, 2004; 2005) apply the EM algorithm for ML estimation of parameters from type-II right-censored and progressively type-II right-censored samples from a bivariate normal distribution, progressively type II censored samples from log normal and Weibull distributions, work out the covariance matrix of estimates using the missing information principle, and determine optimal progressive censored plans for the Weibull distribution using three optimality criteria.

Ng and McLachlan (2003b) have proposed an EM-based semi-parametric mixture model approach to the regression analysis of competing-risks data. With their approach, the component-baseline hazard functions are completely unspecified. More recently, Ng,

McLachlan, Yau, and Lee (2004) considered an EM-based approach to the fitting of mixtures of survival functions with random effects adjustment. The latter allows, for example, for correlations between the survival times of patients from the same hospital.

5.6 EXAMPLE 5.4: CONTINGENCY TABLES WITH INCOMPLETE DATA

In the previous two examples, the ECM algorithm has been implemented with the vector of unknown parameters Ψ partitioned into a number of subvectors, where each CM-step corresponds to maximization over one subvector with the remaining subvectors fixed at their current values. We now give an example where this is not the case. It concerns incomplete-data ML estimation of the cell probabilities of a $2 \times 2 \times 2$ contingency table under a log linear model without the three-way interaction term. It is well known that, in this case, there are no simple formulas for computing the expected cell frequencies or finding the MLE's of the cell probabilities, and an iterative method such as the Iterative Proportional Fitting has to be used; see, for instance, Christensen (1990).

Let θ_{hij} be the probability for the (h, i, j) th cell ($h, i, j = 1, 2$), where the parameter space Ω is the subspace of $\{\theta_{hij}, h, i, j = 1, 2\}$ such that the three-way interaction is zero. Starting from the constant table (that is, $\theta_{hij} = 1/8$), given the fully observed cell events $y = \{y_{hij}\}$ and the current estimates $\theta_{hij}^{(k)}$ of the cell probabilities, the $(k + 1)$ th iteration of Iterative Proportional Fitting is the final output of the following set of three steps:

$$\theta_{hij}^{(k+1/3)} = \theta_{hi(j)}^{(k)} \frac{y_{hi+}}{n}, \quad (5.39)$$

$$\theta_{hij}^{(k+2/3)} = \theta_{h(i)j}^{(k+1/3)} \frac{y_{h+j}}{n}, \quad (5.40)$$

and

$$\theta_{hij}^{(k+3/3)} = \theta_{(h)ij}^{(k+2/3)} \frac{y_{+ij}}{n}, \quad (5.41)$$

where n is the total count,

$$y_{hi+} = \sum_j y_{hij}$$

define the two-way marginal totals for the first two factors,

$$\theta_{hi(j)} = \theta_{hij} / \sum_j \theta_{hij}$$

define the conditional probabilities of the third factor given the first two, etc. It is easy to see that (5.39) corresponds to maximizing the likelihood function subject to the constraints

$$\theta_{hi(j)} = \theta_{hi(j)}^{(k)}$$

for all h, i, j . Similarly, (5.40) and (5.41) correspond to maximizing the likelihood function subject to the constraints

$$\theta_{h(i)j} = \theta_{h(i)j}^{(k+1/3)}$$

and

$$\theta_{(h)ij} = \theta_{(h)ij}^{(k+2/3)},$$

respectively.

Here the notation

$$\theta_{hij}^{(k+s/3)}$$

corresponds to the value of θ_{hij} on the s th of the three CM-steps on the $(k+1)$ th iteration of the ECM algorithm. The simplicity of Iterative Proportional Fitting comes from the fact that the constraint of no three-way iteration only imposes restrictions on the conditional probabilities.

Once each iteration of Iterative Proportional Fitting is identified as a set of conditional maximizations, we can immediately add an E-step at each iteration to develop an algorithm to estimate cell probabilities when data are incomplete. For instance, the only difference between the ECM algorithm and Iterative Proportional Fitting for the above example is to replace y_{ij+} by

$$E_{\Psi^{(k)}}(y_{hi+} \mid \mathbf{y}),$$

with analogous replacements for y_{h+j} and y_{++j} at each iteration. Thus, in this case, as noted by Meng and Rubin (1993), the ECM algorithm can be viewed as a natural generalization of Iterative Proportional Fitting in the presence of incomplete data.

5.7 ECME ALGORITHM

Liu and Rubin (1994) present an extension of their ECM algorithm called the ECME (expectation–conditional maximization either) algorithm. Here the “either” refers to the fact that with this extension, some or all of the CM-steps of the ECM algorithm are replaced by steps that conditionally maximize the incomplete-data log likelihood function, $\log L(\Psi)$, and not the Q -function. Hence with the ECME algorithm, each CM-step either maximizes the conditional expectation of the complete-data log likelihood

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}}\{L_c(\Psi) \mid \mathbf{y}\},$$

or the actual (incomplete-data) log likelihood function, $\log L(\Psi)$, subject to the same constraints on Ψ .

Typically, the ECME algorithm is more tedious to code than the ECM algorithm, but the reward of faster convergence is often worthwhile, especially because it allows convergence to be more easily assessed. As noted previously, the ECM algorithm is an extension of the EM algorithm that typically converges more slowly than the EM algorithm in terms of iterations, but can be much faster in total computer time. Liu and Rubin (1994) find that their extension is nearly always faster than both the EM and ECM algorithms in terms of the number of iterations and moreover can be faster in total computer time by orders of magnitude.

Analogous convergence results hold for the ECME algorithm as for the ECM and EM algorithms. In particular, the ECME algorithm shares with both the EM and ECM algorithms their stable monotone convergence; that is,

$$L(\Psi^{(k+1)}) \geq L(\Psi^{(k)}).$$

As Meng and van Dyk (1997) note, the proof of the above result in Liu and Rubin (1994) contains a technical error and is valid only when all the CM-steps that act on the Q -function are performed *before* those that act on the actual log likelihood, $\log L(\Psi)$.

Liu and Rubin (1994) establish a result on the global speed of convergence of the ECM algorithm analogous to (5.8), which shows that the ECME algorithm typically has a greater

speed of convergence than the ECM algorithm. However, they recognize there are situations where the global speed of convergence of the ECME algorithm is slower than that of the ECM algorithm.

5.8 EXAMPLE 5.5: MAXIMUM LIKELIHOOD ESTIMATION OF t -DISTRIBUTION WITH UNKNOWN DEGREES OF FREEDOM

5.8.1 Application of the EM Algorithm

In Example 2.6 in Section 2.6, we considered the application of the EM algorithm for finding the MLE's of the parameters μ and Σ in the multivariate t -distribution (2.38) with known degrees of freedom ν . We consider here the general case where ν is also unknown, as in Lange et al. (1989). In this more difficult case, Liu and Rubin (1994, 1995) have shown how the MLE's can be found much more efficiently by using the ECME algorithm.

From (2.42) and (2.43), it can be seen that in the general case the E-step on the $(k+1)$ th iteration also requires the calculation of the term

$$E_{\Psi^{(k)}}(\log U_j \mid \mathbf{w}_j) \quad (5.42)$$

for $j = 1, \dots, n$.

To calculate this conditional expectation, we need the result (4.170) that if a random variable R has a gamma (α, β) distribution, then

$$E(\log R) = \psi(\alpha) - \log \beta, \quad (5.43)$$

where

$$\psi(s) = \{\partial \Gamma(s)/\partial s\}/\Gamma(s)$$

is the Digamma function.

Applying the result (5.43) to the conditional density of U_j given w_j , as specified by (2.44), it follows that

$$\begin{aligned} E_{\Psi^{(k)}}(\log U_j \mid \mathbf{w}_j) &= \psi\left(\frac{\nu^{(k)} + p}{2}\right) - \log\left[\frac{1}{2}\{\nu^{(k)} + \delta(\mathbf{w}_j, \boldsymbol{\mu}_j^{(k)}; \boldsymbol{\Sigma}^{(k)})\}\right] \\ &= \log u_j^{(k)} + \{\psi\left(\frac{\nu^{(k)} + p}{2}\right) - \log\left(\frac{\nu^{(k)} + p}{2}\right)\}, \end{aligned} \quad (5.44)$$

where

$$u_j^{(k)} = E_{\Psi^{(k)}}(U_j \mid \mathbf{w}_j) \quad (5.45)$$

$$= \frac{\nu^{(k)} + p}{\nu^{(k)} + \delta(\mathbf{w}_j, \boldsymbol{\mu}^{(k)}; \boldsymbol{\Sigma}^{(k)})} \quad (5.46)$$

for $j = 1, \dots, n$. The last term on the right-hand side of (5.44),

$$\psi\left(\frac{\nu^{(k)} + p}{2}\right) - \log\left(\frac{\nu^{(k)} + p}{2}\right),$$

can be interpreted as the correction for just imputing the mean value $u_j^{(k)}$ for u_j in $\log u_j$.

On using the results (2.47) and (5.44) to calculate the conditional expectation of the complete-data log likelihood from (2.42) and (2.41), we have that $Q(\Psi; \Psi^{(k)})$ is given, on ignoring terms not involving ν , by

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) &= -n \log \Gamma\left(\frac{1}{2}\nu\right) + \frac{1}{2}n\nu \log\left(\frac{1}{2}\nu\right) + \frac{1}{2}n\nu \left\{ \frac{1}{n} \sum_{j=1}^n (\log u_j^{(k)} - u_j^{(k)}) \right. \\ &\quad \left. + \psi\left(\frac{\nu^{(k)} + p}{2}\right) - \log\left(\frac{\nu^{(k)} + p}{2}\right) \right\}. \end{aligned} \quad (5.47)$$

5.8.2 M-Step

On the M-step at the $(k+1)$ th iteration of the EM algorithm with unknown ν , the computation of μ and Σ is the same as that with known ν . On calculating the left-hand side of the equation,

$$\partial Q(\Psi; \Psi^{(k)}) / \partial \nu = 0,$$

it follows that $\nu^{(k+1)}$ is a solution of the equation

$$\begin{aligned} -\psi\left(\frac{1}{2}\nu\right) + \log\left(\frac{1}{2}\nu\right) + 1 + \frac{1}{n} \sum_{j=1}^n (\log u_j^{(k)} - u_j^{(k)}) \\ + \psi\left(\frac{\nu^{(k)} + p}{2}\right) - \log\left(\frac{\nu^{(k)} + p}{2}\right) = 0. \end{aligned} \quad (5.48)$$

5.8.3 Application of ECM Algorithm

Liu and Rubin (1995) note that the convergence of the EM algorithm is slow for unknown ν and the one-dimensional search for the computation of $\nu^{(k+1)}$ is time consuming as discussed and illustrated in Lange et al. (1989). Consequently, they considered extensions of EM that can be more efficient, which we now present.

We consider the ECM algorithm for this problem, where Ψ is partitioned as $(\Psi_1^T, \Psi_2)^T$, with Ψ_1 containing μ and the distinct elements of Σ and with Ψ_2 a scalar equal to ν .

On the $(k+1)$ th iteration of the ECM algorithm, the E-step is the same as given above for the EM algorithm, but the M-step of the latter is replaced by two CM-steps, as follows:

CM-Step 1. Calculate $\Psi_1^{(k+1)}$ by maximizing $Q(\Psi; \Psi^{(k)})$ with Ψ_2 fixed at $\Psi_2^{(k)}$; that is, ν fixed at $\nu^{(k)}$.

CM-Step 2. Calculate $\Psi_2^{(k+1)}$ by maximizing $Q(\Psi; \Psi^{(k)})$ with Ψ_1 fixed at $\Psi_1^{(k+1)}$.

But as $\Psi_1^{(k+1)}$ and $\Psi_2^{(k+1)}$ are calculated independently of each other on the M-step, these two CM-steps of the ECM algorithm are equivalent to the M-step of the EM algorithm. Hence there is no difference between this ECM and the EM algorithms here. But Liu and Rubin (1995) used the ECM algorithm to give two modifications that are different from the EM algorithm. These two modifications are a multicycle version of the ECM algorithm and an ECME extension. The multicycle version of the ECM algorithm has an additional E-step between the two CM-steps. That is, after the first CM-step, the E-step is taken with

$$\begin{aligned} \Psi &= \Psi^{(k+1/2)} \\ &= (\Psi_1^{(k+1)^T}, \Psi_2^{(k)})^T, \end{aligned}$$

instead of with $\Psi = (\Psi_1^{(k)^T}, \Psi_2^{(k)})^T$ as on the commencement of the $(k + 1)$ th iteration of the ECM algorithm.

5.8.4 Application of ECME Algorithm

The ECME algorithm as applied by Liu and Rubin (1995) to this problem is the same as the ECM algorithm, apart from the second CM-step where $\Psi_2 = \nu$ is chosen to maximize the actual likelihood function $L(\Psi)$, as given by (2.39), with Ψ_1 fixed at $\Psi_1^{(k+1)}$.

On fixing Ψ_1 at $\Psi_1^{(k+1)}$, we have from (2.39)

$$\begin{aligned} \log L(\Psi^{(k+1)}) &= -\frac{1}{2}np \log \pi \\ &\quad + n\{\log \Gamma\left(\frac{\nu+p}{2}\right) - \log \Gamma\left(\frac{1}{2}\nu\right)\} \\ &\quad - \frac{1}{2}n \log |\Sigma^{(k+1)}| + \frac{1}{2}n \log \nu \\ &\quad - \frac{1}{2}(\nu+p) \sum_{j=1}^n \log\{\nu + \delta(\mathbf{w}_j, \boldsymbol{\mu}^{(k+1)}; \Sigma^{(k+1)})\}. \end{aligned} \tag{5.49}$$

Thus the second CM-step of the ECME algorithm chooses $\nu^{(k+1)}$ to maximize (5.49) with $\boldsymbol{\mu} = \boldsymbol{\mu}^{(k+1)}$ and $\Sigma = \Sigma^{(k+1)}$. This implies that $\nu^{(k+1)}$ is a solution of the equation

$$\begin{aligned} -\psi\left(\frac{1}{2}\nu\right) + \log\left(\frac{1}{2}\nu\right) + 1 + \frac{1}{n} \sum_{j=1}^n \{\log u_j^{(k+1)}(\nu) - u_j^{(k+1)}(\nu)\} \\ + \psi\left(\frac{\nu+p}{2}\right) - \log\left(\frac{\nu+p}{2}\right) = 0, \end{aligned} \tag{5.50}$$

where

$$u_j^{(k+1)}(\nu) = \frac{\nu+p}{\nu + \delta(\mathbf{w}_j, \boldsymbol{\mu}^{(k+1)}; \Sigma^{(k+1)})}.$$

The solution of this equation involves a one-dimensional search as with the EM algorithm. A comparison of (5.48) with (5.50) demonstrates the difference between the second ECM-step and the second ECME-step in their computation of $\nu^{(k+1)}$.

The multicycle ECM algorithm is obtained by performing an E-step before the second CM-step. The multicycle ECME and ECME algorithms are the same in this application, since the second CM-step of the ECME algorithm is with respect to the actual log likelihood function and not the Q -function.

5.8.5 Some Standard Results

In the next subsection where we are to consider the case of missing data, we need the following results that are well known from multivariate normal theory.

Suppose that a random vector \mathbf{W} is partitioned into two subvectors \mathbf{W}_1 and \mathbf{W}_2 such that

$$\mathbf{W} = (\mathbf{W}_1^T, \mathbf{W}_2^T)^T,$$

where \mathbf{W}_i is of dimension p_i , and $p_1 + p_2 = p$. Further, let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$ and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

be the corresponding partitions of $\boldsymbol{\mu}$ and Σ .

If $\mathbf{W} \sim N(\boldsymbol{\mu}, \Sigma)$, then the conditional distribution of \mathbf{W}_1 given $\mathbf{W}_2 = \mathbf{w}_2$ is p_1 -variate normal with mean

$$\boldsymbol{\mu}_{1:2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{w}_2 - \boldsymbol{\mu}_2) \quad (5.51)$$

and covariance matrix

$$\boldsymbol{\Sigma}_{11:2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \quad (5.52)$$

5.8.6 Missing Data

We now consider the implementation of the EM algorithm and its extensions when some of the \mathbf{w}_j have missing data (assumed to be missing at random; see Section 6.14). For a vector \mathbf{w}_j with missing data, we partition it as

$$\mathbf{w}_j = (\mathbf{w}_{1j}^T, \mathbf{w}_{2j}^T)^T, \quad (5.53)$$

where \mathbf{w}_{2j} is the subvector of dimension p_j containing the p_j elements of \mathbf{w}_j for which the observations are missing. We let the corresponding partition of $\boldsymbol{\mu}$ and Σ be given by

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_{1j}^T, \boldsymbol{\mu}_{2j}^T)^T$$

and

$$\Sigma = \begin{pmatrix} \Sigma_{11;j} & \Sigma_{12;j} \\ \Sigma_{21;j} & \Sigma_{22;j} \end{pmatrix}.$$

In order to implement the E-step, we have to first find the conditional expectations of U_j , $\log U_j$, $U_j \mathbf{W}_j$, and $U_j \mathbf{W}_j \mathbf{W}_j^T$ given the observed data \mathbf{y} (effectively, \mathbf{w}_{1j}), where now \mathbf{y} is given by

$$\mathbf{y} = (\mathbf{w}_{1j}^T, \dots, \mathbf{w}_{nj}^T)^T.$$

The calculation of $u_j^{(k)} = E_{\Psi^{(k)}}(U_j | \mathbf{w}_{1j})$ and $E_{\Psi^{(k)}}(\log U_j | \mathbf{w}_{1j})$ is straightforward in that we simply replace \mathbf{w}_j by its observed subvector \mathbf{w}_{1j} in (5.46) and (5.44), respectively.

To calculate $E_{\Psi^{(k)}}(U_j \mathbf{W}_j | \mathbf{w}_{1j})$, we first take the expectation of $U_j \mathbf{W}_j$ conditional on U_j as well as \mathbf{w}_{1j} and note that the conditional expectation of \mathbf{W}_j does not depend on u_j . This gives

$$\begin{aligned} E_{\Psi^{(k)}}(U_j \mathbf{W}_j | \mathbf{w}_{1j}) &= E_{\Psi^{(k)}}(U_j | \mathbf{w}_{1j}) E_{\Psi^{(k)}}(\mathbf{W}_j | \mathbf{w}_{1j}) \\ &= u_j^{(k)} \mathbf{w}_{2j}^{(k)}, \end{aligned} \quad (5.54)$$

where

$$\begin{aligned} \mathbf{w}_j^{(k)} &= E_{\Psi^{(k)}}(\mathbf{W}_j | \mathbf{w}_{1j}) \\ &= E_{\Psi^{(k)}}(\mathbf{W}_j | \mathbf{w}_{1j}, u_j) \\ &= (\mathbf{w}_{1j}^T, \mathbf{w}_{2j}^{(k)T})^T, \end{aligned} \quad (5.55)$$

and where

$$\begin{aligned} \mathbf{w}_{2j}^{(k)} &= E_{\Psi^{(k)}}(\mathbf{W}_{2j} | \mathbf{w}_{1j}, u_j) \\ &= \boldsymbol{\mu}_{1j}^{(k)} + \boldsymbol{\Sigma}_{12;j}^{(k)} \boldsymbol{\Sigma}_{22;j}^{(k)-1} (\mathbf{w}_{2j}^{(k)} - \boldsymbol{\mu}_{2j}^{(k)}). \end{aligned} \quad (5.56)$$

This last result is obtained on using (5.51).

The expectation in the remaining term $E_{\Psi^{(k)}}(U_j \mathbf{W}_j \mathbf{W}_j^T \mid \mathbf{w}_{1j})$ is also approached by first conditioning on U_j as well as \mathbf{w}_{1j} to give

$$\begin{aligned} E_{\Psi^{(k)}}(U_j \mathbf{W}_j \mathbf{W}_j^T \mid \mathbf{w}_{1j}) &= E_{\Psi^{(k)}}\{U_j E_{\Psi^{(k)}}(\mathbf{W}_j \mathbf{W}_j^T \mid \mathbf{w}_{1j}, U_j)\} \\ &= E_{\Psi^{(k)}}[U_j \{\text{cov}_{\Psi^{(k)}}(\mathbf{W}_j \mid \mathbf{w}_{1j}, U_j) + \mathbf{w}_j(k) \mathbf{w}_j^{(k)^T}\}] \\ &= E_{\Psi^{(k)}}\{U_j \text{cov}_{\Psi^{(k)}}(\mathbf{W}_j \mid \mathbf{w}_{1j}, U_j)\} + \{u_j^{(k)} \mathbf{w}_j^{(k)} \mathbf{w}_j^{(k)^T}\} \\ &= \mathbf{c}_j^{(k)} + u_j^{(k)} \mathbf{w}_j^{(k)} \mathbf{w}_j^{(k)^T}, \end{aligned} \quad (5.57)$$

where

$$\mathbf{c}_j^{(k)} = E_{\Psi^{(k)}}\{U_j \text{cov}_{\Psi^{(k)}}(\mathbf{W}_j \mid \mathbf{w}_{1j}), U_j\}.$$

The (h, i) th element of $\mathbf{c}_j^{(k)}$ is zero if either $(\mathbf{w}_j)_h$ or $(\mathbf{w}_j)_i$ is observed and, if both are missing, it is the corresponding element of

$$\Sigma_{11;j}^{(k)} - \Sigma_{12;j}^{(k)} \Sigma_{22;j}^{(k)^{-1}} \Sigma_{21;j}^{(k)},$$

which is obtained on using (5.52). Alternatively, it can be found numerically by applying the sweep operator to $\boldsymbol{\mu}_j^{(k)}$ and $\Sigma_j^{(k)}$ to predict \mathbf{w}_{2j} by its linear regression on \mathbf{w}_{1j} ; see Goodnight (1979) and Little and Rubin (2002, Chapter 7).

With the conditional expectations calculated as above, we now can compute the updated estimates $\boldsymbol{\mu}^{(k+1)}$ and $\Sigma^{(k+1)}$, of $\boldsymbol{\mu}$ and Σ , respectively. It is not difficult to see that, corresponding to (2.48), $\boldsymbol{\mu}^{(k+1)}$ is given by

$$\boldsymbol{\mu}^{(k+1)} = \sum_{j=1}^n u_j^{(k)} \mathbf{w}_j^{(k)} / \sum_{j=1}^n u_j^{(k)}.$$

That is, we simply impute current conditional expectations for any missing observations in \mathbf{w}_j .

The modification to (2.49) for the updated estimate of Σ is not as straightforward in that it effectively involves imputing the current conditional expectations of cross-product terms like $\mathbf{W}_j \mathbf{W}_j^T$. It can be confirmed that

$$\begin{aligned} \Sigma^{(k+1)} &= n^{-1} \sum_{j=1}^n E_{\Psi^{(k)}}(U_j \mathbf{W}_j \mathbf{W}_j^T \mid \mathbf{w}_{1j}) \\ &\quad - (\sum_{j=1}^n u_j^{(k)} \mathbf{w}_j^{(k)}) (\sum_{j=1}^n u_j^{(k)} \mathbf{w}_j^{(k)})^T / \sum_{j=1}^n u_j^{(k)} \\ &= n^{-1} \sum_{j=1}^n \{u_j^{(k)} (\mathbf{w}_j^{(k)} - \boldsymbol{\mu}_j^{(k+1)}) (\mathbf{w}_j^{(k)} - \boldsymbol{\mu}_j^{(k+1)})^T + \mathbf{c}_j^{(k)}\}, \end{aligned} \quad (5.58)$$

on using (5.57).

Similarly, as above, the E- and CM-steps of the ECM and ECME algorithms can be implemented for this problem when there are missing data.

5.8.7 Numerical Examples

We report here first the numerical example of Liu and Rubin (1994) who consider the data set in Table 1 of Cohen, Dalal, and Tukey (1993), which contains 79 bivariate observations $w_j (j = 1, \dots, 79)$. Liu and Rubin (1994) create missing data by deleting the second component of every third observation starting from $j = 1$; that is, $j = 1, 4, 7, \dots$, and the first component of every third observation starting from $j = 2$; that is, $j = 2, 5, 8, \dots$, and treating the deleted observations as being missing at random. Liu and Rubin (1994) fitted the t -distribution to this missing data set via the EM (=ECM) and ECME algorithms, starting each from $\nu^{(0)} = 1,000$ and $\mu^{(0)}$ and $\Sigma^{(0)}$ equal to their MLE's under the bivariate normal model. The values of $\log L(\Psi^{(k)})$ and $\log \nu^{(k)}$ corresponding to each of these algorithms are displayed in Figure 5.1, where the solid and dashed lines represent the values from the EM and ECME algorithms, respectively. The dramatically faster convergence of the ECME algorithm over the EM algorithm is obvious, and is so dramatic that the fact that the starting values are the same is lost in the plots, which display the results as functions of $\log k$.

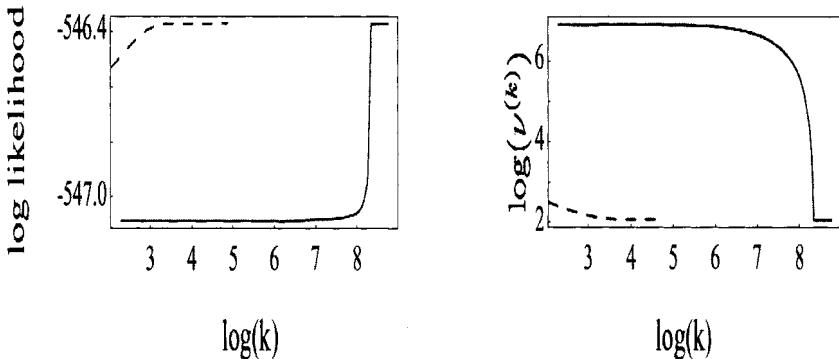


Figure 5.1 Multivariate t : Convergence of EM (solid line) and ECME (dashed line) for log likelihood and $\log \nu^{(k)}$. From Liu and Rubin (1994).

In another example, Liu and Rubin (1995) compare the performances of the ECME, EM, and multicycle EM algorithms in their application to an artificial data set created by appending four extreme observations to the artificial bivariate data set of Murray (1977), given in Section 3.6.1. They find that the speed of convergence of the ECME algorithm is significantly faster and, in terms of the actual computational time, to have a seven-fold advantage over the other methods. In later work to be considered shortly in Section 5.12, Meng and van Dyk (1997) report results in which the ECME algorithm offers little advantage over the multicycle EM algorithm for this problem.

5.8.8 Theoretical Results on the Rate of Convergence

Liu and Rubin (1994) provide explicit large-sample results for the rate of convergence (equivalently, the speed of convergence) for ML estimation of the parameters of the univariate t -distribution. They show that the ECME algorithm has a smaller rate of convergence

for this problem than the EM algorithm. In the case of known degrees of freedom ν , they note that the large-sample rate of convergence is $3/(3 + \nu)$, which was obtained also by Dempster et al. (1980); and Rubin (1994) give similar calculations for a contamination model.

5.9 EXAMPLE 5.6: VARIANCE COMPONENTS

5.9.1 A Variance Components Model

We consider the following variance components, or repeated measures model, as considered previously by Harville (1977), Laird and Ware (1982), Laird, Lange, and Stram (1987), and Liu and Rubin (1994). Specifically, we let \mathbf{y}_j denote the $n_j \times 1$ vector of n_j measurements observed on the j th experimental unit, where it is assumed that

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \mathbf{e}_j \quad (j = 1, \dots, m). \quad (5.59)$$

Here \mathbf{X}_j and \mathbf{Z}_j are known $n_j \times p$ and $n_j \times q$ design matrices, respectively, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects to be estimated, and $n = \sum_{j=1}^m n_j$. The $q \times 1$ random effects vector \mathbf{b}_j is distributed $N(\mathbf{0}, D)$, independently of the error vector \mathbf{e}_j , which is distributed

$$N(\mathbf{0}, \sigma^2 \mathbf{R}_j),$$

where \mathbf{R}_j is a known $n_j \times n_j$ positive definite symmetric matrix and where D is an unknown $q \times q$ positive definite symmetric matrix of parameters to be estimated along with the unknown positive scalar σ^2 ($j = 1, \dots, m$). It is assumed further that \mathbf{b}_j and \mathbf{e}_j are distributed independently for $j = 1, \dots, m$.

For this problem the observed data vector is $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_m^T)^T$, for which the incomplete-data log likelihood, $\log L(\Psi)$, is given by

$$\begin{aligned} \log L(\Psi) &= -\frac{1}{2} \sum_{j=1}^m \log |\Sigma_{11;j}| \\ &\quad -\frac{1}{2} \sum_{j=1}^m (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta})^T \Sigma_{11;j}^{-1} (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}), \end{aligned} \quad (5.60)$$

where

$$\Sigma_{11;j} = \mathbf{Z}_j D \mathbf{Z}_j^T + \sigma^2 \mathbf{R}_j \quad (j = 1, \dots, m).$$

An obvious choice for the complete-data vector is

$$\mathbf{x} = (\mathbf{x}_1^T, \dots, \mathbf{x}_m^T)^T$$

where

$$\mathbf{x}_j = (\mathbf{y}_j^T, \mathbf{b}_j^T)^T \quad (j = 1, \dots, m);$$

that is,

$$\mathbf{z} = (\mathbf{b}_1^T, \dots, \mathbf{b}_m^T)^T$$

is the missing-data vector. Under the assumptions of the model (5.59), it follows that \mathbf{x}_j has a multivariate normal distribution with mean

$$\boldsymbol{\mu}_j = \begin{pmatrix} \mathbf{X}_j\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix} \quad (5.61)$$

and covariance matrix

$$\Sigma_j = \begin{pmatrix} \mathbf{Z}_j \mathbf{D} \mathbf{Z}_j^T + \sigma^2 \mathbf{R}_j & \mathbf{Z}_j \mathbf{D} \\ \mathbf{D} \mathbf{Z}_j^T & \mathbf{D} \end{pmatrix} \quad (5.62)$$

for $j = 1, \dots, m$. Note that here \mathbf{X}_j and \mathbf{Z}_j are design matrices and are not random matrices (vectors) corresponding to \mathbf{x}_j and \mathbf{z}_j , respectively.

The vector Ψ of unknown parameters is given by the elements of β , σ^2 , and the distinct elements of \mathbf{D} . From (5.61) and (5.62), the complete-data log likelihood is given, apart from an additive constant, by

$$\begin{aligned} \log L_c(\Psi) &= \sum_{j=1}^m \log \phi(\mathbf{x}_j; \boldsymbol{\mu}_j, \Sigma_j) \\ &= -\frac{1}{2} \sum_{j=1}^m \{ \log |\Sigma_j| + (\mathbf{x}_j - \boldsymbol{\mu}_j)^T \Sigma_j^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_j) \}. \end{aligned} \quad (5.63)$$

5.9.2 E-Step

Considering the E-step of the EM algorithm, it can be seen from (5.63), that in order to compute the conditional expectation of (5.63) given the observed data \mathbf{y} , we require the following conditional moments of the missing random effects vector \mathbf{b}_j , namely

$$E_{\Psi^{(k)}}(\mathbf{b}_j | \mathbf{y}_j)$$

and

$$E_{\Psi^{(k)}}(\mathbf{b}_j \mathbf{b}_j^T | \mathbf{y}_j).$$

These are directly obtainable from the well-known results in multivariate theory given by (5.51) and (5.52).

From these results applied to the joint distribution of \mathbf{y}_j and \mathbf{b}_j , we have that the conditional distribution of \mathbf{b}_j given \mathbf{y}_j is multivariate normal with mean

$$\begin{aligned} \boldsymbol{\mu}_{2 \cdot 1; j} &= \mathbf{D} \mathbf{Z}_j^T (\mathbf{Z}_j \mathbf{D} \mathbf{Z}_j^T + \sigma^2 \mathbf{R}_j)^{-1} (\mathbf{y}_j - \mathbf{X}_j \beta) \\ &= \mathbf{D} \mathbf{Z}_j^T (\mathbf{R}_j^{-1} \mathbf{Z}_j \mathbf{D} \mathbf{Z}_j^T + \sigma^2 \mathbf{I}_{n_j})^{-1} \mathbf{R}_j^{-1} (\mathbf{y}_j - \mathbf{X}_j \beta) \end{aligned} \quad (5.64)$$

and covariance matrix

$$\Sigma_{22 \cdot 1; j} = \mathbf{D} - \mathbf{D} \mathbf{Z}_j^T (\mathbf{Z}_j \mathbf{D} \mathbf{Z}_j^T + \sigma^2 \mathbf{R}_j)^{-1} \mathbf{Z}_j \mathbf{D}$$

for $j = 1, \dots, m$.

To show the close connection between least-squares computation and (5.64) and (5.65), we make use of the identity

$$(\mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \sigma^2 \mathbf{D}^{-1}) \mathbf{D} \mathbf{Z}_j^T = \mathbf{Z}_j^T \mathbf{R}_j^{-1} (\mathbf{Z}_j \mathbf{D} \mathbf{Z}_j^T + \sigma^2 \mathbf{R}_j),$$

from which

$$\boldsymbol{\mu}_{2 \cdot 1; j} = (\mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \sigma^2 \mathbf{D}^{-1})^{-1} \mathbf{Z}_j^T \mathbf{R}_j^{-1} (\mathbf{y}_j - \mathbf{X}_j \beta), \quad (5.65)$$

and

$$\begin{aligned} \Sigma_{22 \cdot 1; j} &= \mathbf{D} - (\mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \sigma^2 \mathbf{D}^{-1})^{-1} \mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j \mathbf{D} \\ &= (\mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \sigma^2 \mathbf{D}^{-1})^{-1} \{ (\mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \sigma^2 \mathbf{D}^{-1}) \mathbf{D} - \mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j \mathbf{D} \} \\ &= (\sigma^{-2} \mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \mathbf{D}^{-1})^{-1}. \end{aligned} \quad (5.66)$$

From (5.65) and (5.66), we have that

$$E_{\Psi^{(k)}}(\mathbf{b}_j \mid \mathbf{y}_j) = \mathbf{b}_j^{(k)},$$

where

$$\mathbf{b}_j^{(k)} = (\mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \sigma^{(k)^2} \mathbf{D}^{(k)^{-1}})^{-1} \mathbf{Z}_j^T \mathbf{R}_j^{-1} (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta}^{(k)}) \quad (5.67)$$

and

$$\begin{aligned} E_{\Psi^{(k)}}(\mathbf{b}_j \mathbf{b}_j^T \mid \mathbf{y}_j) &= \text{cov}_{\Psi^{(k)}}(\mathbf{b}_j \mid \mathbf{y}_j) + \mathbf{b}_j^{(k)} \mathbf{b}_j^{(k)T} \\ &= (\sigma^{(k)^{-2}} \mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \mathbf{D}^{(k)^{-1}})^{-1} + \mathbf{b}_j^{(k)} \mathbf{b}_j^{(k)T}. \end{aligned} \quad (5.68)$$

5.9.3 M-Step

On the M-step at the $(k+1)$ th iteration, we have to choose $\boldsymbol{\beta}^{(k+1)}$, $\sigma^{(k+1)^2}$, and $\mathbf{D}^{(k+1)}$ to maximize $Q(\Psi; \Psi^{(k)})$. Now

$$\boldsymbol{\beta}^{(k+1)} = (\sum_{j=1}^m \mathbf{X}_j^T \mathbf{R}_j^{-1} \mathbf{X}_j)^{-1} \sum_{j=1}^m \mathbf{X}_j^T \mathbf{R}_j^{-1} (\mathbf{y}_j - \mathbf{Z}_j \mathbf{b}_j^{(k)})$$

The iterates $\mathbf{D}^{(k+1)}$ and $\sigma^{(k+1)^2}$ can be obtained either by direct differentiation of $Q(\Psi; \Psi^{(k)})$ or by noting that $L_c(\Psi)$ belongs to the exponential family and considering the conditional expectations of the complete-data sufficient statistics for $\boldsymbol{\beta}$, \mathbf{D} , and σ^2 .

It follows that

$$\mathbf{D}^{(k+1)} = \frac{1}{m} \sum_{j=1}^m E_{\Psi^{(k)}}\{\mathbf{b}_j \mathbf{b}_j^T \mid \mathbf{y}_j\} \quad (5.69)$$

and

$$\sigma^{(k+1)^2} = \frac{1}{n} \sum_{j=1}^m E_{\Psi^{(k)}}\{\mathbf{e}_j^T \mathbf{R}_j^{-1} \mathbf{e}_j \mid \mathbf{y}_j\}.$$

Now

$$\begin{aligned} E_{\Psi^{(k)}}\{\mathbf{e}_j^T \mathbf{R}_j^{-1} \mathbf{e}_j \mid \mathbf{y}_j\} &= \text{tr} \mathbf{R}_j^{-1} E_{\Psi^{(k)}}\{\mathbf{e}_j \mathbf{e}_j^T \mid \mathbf{y}_j\} \\ &= \text{tr} \mathbf{R}_j^{-1} [\text{cov}_{\Psi^{(k)}}\{\mathbf{e}_j \mid \mathbf{y}_j\} + \mathbf{e}_j^{(k)} \mathbf{e}_j^{(k)T}] \\ &= \text{tr}[\mathbf{R}_j^{-1} \text{cov}_{\Psi^{(k)}}\{\mathbf{e}_j \mid \mathbf{y}_j\}] + \mathbf{e}_j^{(k)T} \mathbf{R}_j^{-1} \mathbf{e}_j^{(k)}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{e}_j^{(k)} &= E_{\Psi^{(k)}}(\mathbf{e}_j \mid \mathbf{y}_j) \\ &= \mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta}^{(k)} - \mathbf{Z}_j \mathbf{b}_j^{(k)}, \end{aligned}$$

and where

$$\begin{aligned} \text{cov}_{\Psi^{(k)}}(\mathbf{e}_j \mid \mathbf{y}_j) &= \text{cov}_{\Psi^{(k)}}(\mathbf{Z}_j \mathbf{b}_j \mid \mathbf{y}_j) \\ &= \mathbf{Z}_j (\sigma^{(k)^{-2}} \mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \mathbf{D}^{(k)^{-1}})^{-1} \mathbf{Z}_j^T. \end{aligned}$$

Thus

$$\begin{aligned} E_{\Psi^{(k)}}(\mathbf{e}_j^T \mathbf{R}_j^{-1} \mathbf{e}_j \mid \mathbf{y}_j) &= \text{tr}\{\mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j (\sigma^{(k)^{-2}} \mathbf{Z}_j^T \mathbf{R}_j^{-1} \mathbf{Z}_j + \mathbf{D}^{(k)^{-1}})\} \\ &\quad + \mathbf{e}_j^{(k)T} \mathbf{R}_j^{-1} \mathbf{e}_j^{(k)}. \end{aligned}$$

5.9.4 Application of Two Versions of ECME Algorithm

For the variance components model (5.59), Liu and Rubin (1994) consider two versions of the ECME algorithm that can be especially appropriate when the dimension q of \mathbf{D} is relatively large. In Version 1 of the ECME algorithm, Ψ is partitioned as $(\Psi_1^T, \Psi_2^T)^T$, where Ψ_1 contains the distinct elements of \mathbf{D} and σ^2 and $\Psi_2 = \beta$. The two CM-steps are as follows.

CM-Step 1. Calculate $\mathbf{D}^{(k+1)}$ and $\sigma^{(k+1)^2}$ as above on the M-step of the EM algorithm.

CM-Step 2. Calculate $\beta^{(k+1)}$ as

$$\beta^{(k+1)} = \left\{ \sum_{j=1}^m \mathbf{X}_j^T \Sigma_{11;j}^{(k+1)^{-1}} \mathbf{X}_j \right\}^{-1} \left\{ \sum_{j=1}^m \mathbf{X}_j^T \Sigma_{11;j}^{(k+1)^{-1}} \mathbf{y}_j \right\},$$

where

$$\Sigma_{11;j}^{(k+1)} = \mathbf{Z}_j \mathbf{D}^{(k+1)} \mathbf{Z}_j^T + \sigma^{(k+1)^2} \mathbf{R}_j \quad (j = 1, \dots, m),$$

which is the value of β that maximizes the incomplete-data log likelihood $\log L(\Psi)$ over β with $\mathbf{D} = \mathbf{D}^{(k+1)}$ and $\sigma^2 = \sigma^{(k+1)^2}$.

As pointed out by Liu and Rubin (1994), this corresponds to the algorithm given by Laird and Ware (1982) and Laird et al. (1987), who mistakenly called it an EM algorithm. The algorithm was called a hybrid EM algorithm by Jennrich and Schluchter (1986), who also realized it is not an EM algorithm.

In Version 2 of the ECME algorithm proposed in Liu and Rubin (1994), Ψ is partitioned as $(\Psi_1^T, \Psi_2^T, \Psi_3)^T$, where Ψ_1 consists of the distinct elements of \mathbf{D} , $\Psi_2 = \beta$, and $\Psi_3 = \sigma^2$. The three CM-steps are as follows.

CM-Step 1. Calculate $\mathbf{D}^{(k+1)}$ as on the M-step of EM.

CM-Step 2. Calculate $\hat{\beta}$ as in the CM-Step 2 above of Version 1 of the ECME algorithm.

CM-Step 3. Calculate $\sigma^{(k+1)^2}$ to maximize $L(\Psi)$ over σ^2 with $\mathbf{D} = \mathbf{D}^{(k+1)}$ and $\beta = \beta^{(k+1)}$. The solution to this step does not exist in closed form and so must be calculated using, say, a quasi-Newton method.

5.9.5 Numerical Example

As a numerical example of the above model, we report the example considered by Liu and Rubin (1994). They analyzed the data of Besag (1991) on the yields of 62 varieties of winter wheat in three physically separated complete replicates, using the first-difference Gaussian noise model for the random fertility effects. Hence in the notation of the model (5.60), $m = 3$, $n_j = 61$, and $n = 183$, where \mathbf{y}_j gives the 61 differences of the yields in replicate j , where \mathbf{X}_j is the corresponding (61×62) design matrix, \mathbf{Z}_j is the (61×61) identity matrix, $\mathbf{R}_j = \mathbf{R}$, the (61×61) Toeplitz matrix with the first row $(2, -1, 0, \dots, 0)$, and where β is the vector of the fixed variety effects with the constraint

$$\sum_{i=1}^6 \beta_i = 0.$$

The covariance matrix \mathbf{D} of the random effects is restricted to have the form

$$\mathbf{D} = \alpha^2 \mathbf{I}_q, \quad (5.70)$$

where α is a nonnegative unknown scalar. The vector of unknown parameters is therefore

$$\boldsymbol{\Psi} = (\boldsymbol{\beta}^T, \alpha^2, \sigma^2)^T.$$

Under the restriction (5.70), the equation (5.68) for $\mathbf{D}^{(k+1)}$ on the $(k+1)$ th iteration of the M-step has to be replaced by the following equation for $\alpha^{(k+1)^2}$,

$$\begin{aligned} \alpha^{(k+1)^2} &= \frac{1}{n} \sum_{j=1}^m E_{\boldsymbol{\Psi}^{(k)}} \{ \mathbf{b}_j^T \mathbf{b}_j \mid \mathbf{y} \} \\ &= \frac{1}{n} \left[\sum_{j=1}^m \mathbf{b}_j^{(k)T} \mathbf{b}_j^{(k)} + \text{tr}\{(\mathbf{R} \sigma^{(k)2} + \alpha^{(k)2} \mathbf{I}_q)^{-1} \sigma^{(k)2} \alpha^{(k)2} \mathbf{R}\} \right]. \end{aligned}$$

Liu and Rubin (1994) applied the EM algorithm and Versions 1 and 2 above of the ECME algorithm with the common starting value

$$\boldsymbol{\Psi}^{(0)} = (\boldsymbol{\beta}^{(0)T}, \sigma^{(0)2}, \alpha^{(0)2})^T,$$

where $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, $\sigma^{(0)2} = \alpha^{(0)2} \times 10^{-4}$, and

$$\alpha^{(0)2} = \frac{1}{m} \sum_{j=1}^m \mathbf{y}_j^T \mathbf{y}_j.$$

The corresponding values of the log likelihood and α are given in Figure 5.2. In the left-side figure, the solid line corresponds to the EM algorithm and the dashed line to Version 1 of the ECME algorithm. In the right-side figure, the dashed line corresponds to Version 1 of the ECME algorithm and the dotted line to Version 2, using an expanded scale. It can be seen from these figures that, in terms of number of iterations, Version 2 is substantially faster than Version 1 of the ECME algorithm, which is substantially faster than the EM algorithm. In terms of total computer time in this example, Liu and Rubin (1994) find that Version 2 of the ECME algorithm is approximately twice as fast as the EM algorithm, which is approximately an order of magnitude faster than Version 1 of the ECME algorithm.

5.10 LINEAR MIXED MODELS

5.10.1 Introduction

In the variance component version of the linear mixed model we discussed in the last section, the likelihood is a function of the regression coefficients of the fixed effects and the covariance matrix of the random effects including the error term. The ECME version of the EM algorithm came in handy for ML estimation upon formulation of the random effects as missing values and led to alternate maximization over $\boldsymbol{\beta}$ given \mathbf{D} and σ^2 and over \mathbf{D} and σ^2 given $\boldsymbol{\beta}$ until convergence. We consider below a more general model than the variance component model of the previous section, called the Linear Mixed Model.

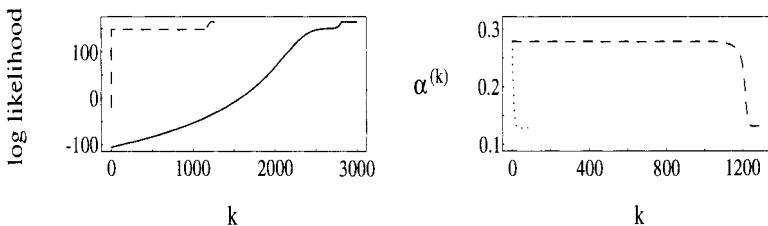


Figure 5.2 Variance Components: Convergence of EM (solid line) and ECME (dashed line) for log likelihood and ECME-1 (dashed line) and ECME-2 (dotted line) for $\alpha^{(k)}$. From Liu and Rubin (1994).

5.10.2 General Form of Linear Mixed Model

We now consider a form more general than (5.59) for the linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}$$

where \mathbf{y} is an n -vector of observations, $\boldsymbol{\beta}$ is a p -vector of fixed effects parameters, \mathbf{X} is an $n \times p$ fixed effects regressor matrix, \mathbf{b} is a q -vector of random effects, \mathbf{Z} is an $n \times q$ random effects regressor matrix, \mathbf{e} is an n -vector of random errors and n is the number of observations. The variables \mathbf{b} and \mathbf{e} are uncorrelated and distributed as

$$\mathbf{b} \sim N_q(\mathbf{0}, \mathbf{D}); \quad \mathbf{e} \sim N_n(\mathbf{0}, \mathbf{R}).$$

Now \mathbf{V} , the covariance matrix of observations, is

$$\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R}.$$

If \mathbf{V} is known, then an estimate of $\boldsymbol{\beta}$ can be obtained by Generalized Least-Squares by minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

But \mathbf{V} is unknown. Hence more complex estimation methods are needed. The EM algorithm is a way of finding MLE in this case; see Pawitan (2001) for details.

ML estimates obtained by this procedure or otherwise are known to be biased in the same way as the sample variance (as MLE with a denominator of n , the sample size) is biased, where a standard way of correcting this bias is to maximize the likelihood based on the deviations from the sample mean (to get a denominator of $n - 1$). Such a procedure could also be adopted for mixed models, where the residuals will be deviations from the fixed part of the model. Such a likelihood is called the Residual, Restricted, or Reduced Likelihood and the procedure of parameter estimation by maximizing this is called the Residual Maximum Likelihood (REML) method. In the Residual Likelihood, the $\boldsymbol{\beta}$ parameters do not appear since the residuals are deviations from the fixed effects; thus, the REML method only gives estimates of variance and covariance parameters. In this context the fixed regression coefficients can be estimated by ordinary least squares using the REML estimates of the variances and covariances, in an iterative scheme. This is the gist of the REML estimation in mixed models. The random effect parameters can be predicted using the fixed effects estimates thus obtained; they are called Best Linear Unbiased Predictors (BLUP's).

5.10.3 REML Estimation

The full likelihood is

$$L(\boldsymbol{\beta}, \mathbf{G}, \mathbf{R}) = (2\pi)^{n/2} |\mathbf{V}|^{-1/2} \exp\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}.$$

If $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X} \mathbf{V}^{-1} \mathbf{y}$$

is the GLS estimator. With $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, the log likelihood $\log L_R$ formed on the basis of \mathbf{e} is called the restricted or residual log likelihood function. It is given by

$$-2 \log L_R = \log \mathbf{V} + \log |\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}| + \mathbf{e}' \mathbf{V}^{-1} \mathbf{e} + \frac{n-p}{2} \log 2\pi,$$

when the rank of \mathbf{X} is p . Another way to consider this Restricted Likelihood is to take a Bayesian viewpoint and treat it as an integrated form of the full likelihood with respect to a locally uniform prior distribution on $\boldsymbol{\beta}$,

$$L(\mathbf{G}, \mathbf{R}) = \int L(\boldsymbol{\beta}, \mathbf{G}, \mathbf{R}) d\boldsymbol{\beta}.$$

The REML estimates are unbiased and have an asymptotic normal distribution. Asymptotically, ML and REML estimates and covariances coincide. REML can also be derived using a Bayes approach under noninformative (uniform) prior for $\boldsymbol{\beta}$ as suggested above. In general, iterative procedures are used to maximize likelihood or restricted likelihood. Hocking (2003) discusses in detail ML and REML methods of estimation in balanced and unbalanced mixed models and develops the EM algorithm in this context.

Generally EM steps are generally quickly and easily implemented and although they rapidly bring the parameters in the optimal region, progress towards optimum is often slow. The Newton-Raphson algorithm on the other hand, though often tedious to compute, can converge rapidly near the optimum. Thus often, especially in the context of mixed models, ML and REML algorithms are implemented by a hybrid of EM and Newton-Raphson algorithms, by initially performing a certain number of EM iterations and then switching over to Newton-Raphson iterations. Bates and DebRoy (2004) provide some heuristics for obtaining starting values for the parameters; see Pinheiro and Bates (2000) and Bates and DebRoy (2004) for more details. Such algorithms are implemented in many standard statistical software packages. In the next subsection, we give an example of an REML computation.

van Dyk (2000) presents efficient algorithms for ML, REML estimation and posterior mode finding in the context of mixed models, combining parameter expansion (PX-EM), conditional data augmentation, and the ECME algorithm.

Foulley, Jaffrezic, and Robert-Granie (2000) discuss REML estimation of covariance components in longitudinal data analysis in Gaussian models using the EM algorithm. In particular, they consider random coefficient models, stationary time processes, and measurement error models.

5.10.4 Example 5.7: REML Estimation in a Hierarchical Random Effects Model

The data below are part of a data set of a large study of the intestinal health in newborn piglets. The data set consists of log transformed measurements of the gut enzyme lactase

in 20 piglets taken from 5 different litters. For each of the 20 piglets the lactase level was measured in three different regions and at the time the measurement was taken the piglet was either unborn (status=1) or newborn (status=2). These data are from the source mentioned at the foot of the table. The number of piglets in the five litters are varied—4, 4, 3, 2, and 7. The data set has one missing value; it has 59 complete observations on the 20 piglets. The case with missing data has been omitted from analysis. The data set is unbalanced.

Table 5.1 Data on piglet lactase.

Litter	Pig	Status	Log(lactase) at 3 regions		
1	1	1	1.89537	1.97046	1.78255
1	2	1	2.24496	1.43413	2.16905
1	3	2	1.74222	1.84277	0.17479
1	4	2	2.12704	1.90954	1.49492
2	5	2	1.62897	2.26642	1.96763
2	6	2	2.01948	2.56443	1.16387
2	7	2	2.20681	2.55652	1.69358
2	8	2	1.09186	1.93091	.
3	9	1	2.36462	2.72261	2.80336
3	10	1	2.42834	2.64971	2.54788
3	11	2	1.58104	1.52606	1.65058
4	12	1	1.97162	2.11342	2.51278
4	13	1	2.06739	2.25631	1.79251
5	14	1	1.93274	1.82394	1.23629
5	15	2	2.07386	1.96713	0.47971
5	16	2	2.01307	1.85483	2.18274
5	17	2	2.86629	2.71414	1.60533
5	18	2	1.97865	1.93342	0.74943
5	19	2	2.89886	2.88606	2.20697
5	20	2	1.87733	1.70260	1.11077

Source: Charlotte Reinhard Bjørnved, Faculty of Life Sciences, University of Copenhagen, Denmark, and taken from the course notes of Per Bruun Brockhoff, with permission from Charlotte Reinhard Bjørnved and Per Bruun Brockhoff.

The data set is hierarchical with litter as the first level, pig as the second and individual measurements as the third. The model is of the form,

$$\log(\text{lactase}) = \text{intercept} + \text{status} + \text{region} + \text{status} * \text{region} + \text{litter} + \text{pig} + \text{error},$$

where status, region, and status*region are fixed effects and litter and pig are random effects and may be suitable for the data. There are three variance components in the model— σ_a^2 , σ_b^2 , and σ_e^2 , respectively of litter, pig and error. The structure of the 59×59 covariance matrix \mathbf{V} of the observations is a five diagonal blocks (for the five litters) of orders 12×12 , 11×11 , 9×9 , 6×6 , and 21×21 , with all off-diagonal blocks consisting of matrices with all elements 0. The covariance matrix of litter 4, for instance is:

$$\left[\begin{array}{cccccc} \sigma_a^2 + \sigma_b^2 + \sigma_e^2 & \sigma_a^2 + \sigma_b^2 & \sigma_a^2 + \sigma_b^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 + \sigma_b^2 & \sigma_a^2 + \sigma_b^2 + \sigma_e^2 & \sigma_a^2 + \sigma_b^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 + \sigma_b^2 & \sigma_a^2 + \sigma_b^2 & \sigma_a^2 + \sigma_b^2 + \sigma_e^2 & \sigma_a^2 \sigma_b^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_b^2 + \sigma_e^2 & \sigma_a^2 + \sigma_b^2 & \sigma_a^2 + \sigma_b^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_b^2 & \sigma_a^2 + \sigma_b^2 + \sigma_e^2 & \sigma_a^2 + \sigma_b^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_b^2 & \sigma_a^2 + \sigma_b^2 & \sigma_a^2 + \sigma_b^2 + \sigma_e^2 \end{array} \right]$$

Other blocks are similar in nature. The REML estimation here is a hybrid ECME (β and V) and Newton-Raphson to produce estimates for fixed effects and these three variance components. After five ECME iterations, we switched over to Newton-Raphson (NR) and convergence was achieved in 25 iterations. Values of $-2 \log L$ are plotted against iteration numbers with ECME or NR symbols are plotted in Figure 5.3. Estimates of the fixed effects and of variance components are provided by the REML method, which in this case are as follows:

$$\hat{\sigma}_a^2 = 0.000; \quad \hat{\sigma}_b^2 = 0.101; \quad \hat{\sigma}_e^2 = 0.121.$$

Fixed Effects Estimates:

Intercept: 1.344; Status: 0.776; Region 1: 0.664; Region 2: 0.783;

Status * Region 1*1: -0.655; 1*2: -0.765.

In a standard analysis, standard errors and tests of significance results, confidence intervals, predictions of random effects (Best Linear Unbiased Predictors—BLUPs) with standard errors and confidence intervals are also provided.

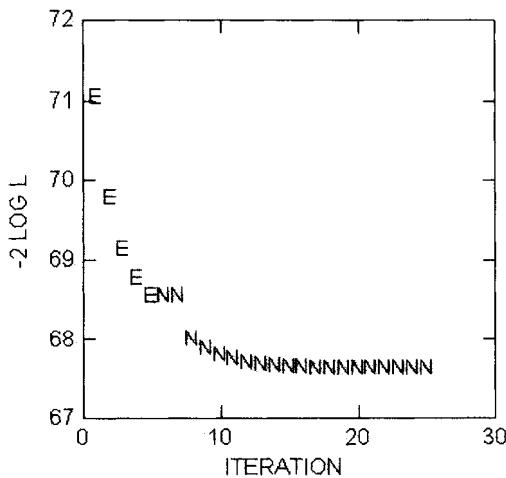


Figure 5.3 REML convergence for piglactase data; E: ECME; N: Newton-Raphson.

5.10.5 Some Other EM-Related Approaches to Mixed Model Estimation

Bates and Pinheiro (1998), Bates and DebRoy (2004), and Pinheiro and Bates (2000) discuss profile log likelihood, profiled restricted log likelihood, penalized least squares and other approaches, some of which include EM iterations. Many books dealing with mixed models and variance-components analysis discuss the use of the EM algorithm, its variations, and EM-hybrid algorithms for ML and REML estimation of mixed model parameters.

Pawitan (2001) derives the EM algorithm for general mixed model estimation. Verbeke and Molenberghs (2000) discuss issues regarding use of the EM algorithm in mixed model estimation. Demidenko (2004) derives the EM algorithm and other algorithms for a general linear mixed model. Recently, in the context of clustering, Ng, McLachlan, Wang, Ben-Tovim Jones, and Ng (2006) have considered the fitting of mixtures of linear mixed models which allow not only for correlations between repeated measurements on an entity, but also between measurements on different entities that belong to the same cluster. They show that the EM algorithm can be implemented exactly for this mixture of linear mixed models. Their procedure is called EMMIX-WIRE (EM-based MIXture analysis WIth Random Effects).

With the EMMIX-WIRE procedure, the observed p -dimensional vectors $\mathbf{y}_1, \dots, \mathbf{y}_n$ are assumed to have come from a mixture of a finite number, say g , of components in some unknown proportions π_1, \dots, π_g , which sum to one. Conditional on its membership of the i th component of the mixture, the distribution of \mathbf{y}_j follows the model

$$\mathbf{y}_j = \mathbf{X}\boldsymbol{\beta}_i + \mathbf{U}\mathbf{b}_{ij} + \mathbf{V}\mathbf{c}_i + \boldsymbol{\epsilon}_{ij} \quad (j = 1, \dots, n), \quad (5.71)$$

where the p_β elements of $\boldsymbol{\beta}_i$ are fixed effects modeling the conditional mean of \mathbf{y}_j in the i th component and where \mathbf{b}_{ij} and \mathbf{c}_i are vectors of random effects of dimension p_b and p_c , respectively ($i = 1, \dots, g$); \mathbf{X} , \mathbf{U} , and \mathbf{V} are known design matrices for the corresponding fixed and random effects. In (5.71), the random effects $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \dots, \mathbf{b}_{ip}^T)^T$ and \mathbf{c}_i , and the measurement error vector $(\boldsymbol{\epsilon}_{i1}^T, \dots, \boldsymbol{\epsilon}_{in}^T)^T$, are assumed to be mutually independent. The distributions of \mathbf{b}_{ij} and \mathbf{c}_i are taken to be multivariate normal $N_{p_b}(\mathbf{0}, \mathbf{H}_i)$ and $N_{p_c}(\mathbf{0}, \theta_{ci} \mathbf{I}_{p_c})$, respectively, where \mathbf{H}_i is a $p_b \times p_b$ covariance matrix. The measurement error vector $\boldsymbol{\epsilon}_{ij}$ is also taken to be multivariate normal $N_p(\mathbf{0}, \mathbf{A}_i)$, where $\mathbf{A}_i = \text{diag}(\mathbf{W}\boldsymbol{\xi}_i)$ is a diagonal matrix constructed from the vector $(\mathbf{W}\boldsymbol{\xi}_i)$ with $\boldsymbol{\xi}_i = (\sigma_{i1}^2, \dots, \sigma_{ip_e}^2)^T$ and \mathbf{W} a known $p \times p_e$ zero-one design matrix. The presence of the random-effects term \mathbf{c}_i in (5.71) induces a correlation between observations from the same component. Ng et al. (2006) show how the E-step for this model can be implemented exactly.

5.10.6 Generalized Linear Mixed Models

In the generalized linear mixed model (GLMM), the EM algorithm gives rise to difficulties because the E-step even under Gaussian assumptions of random effects is intractable. We discuss this briefly in Section 6.3.4 in the next chapter. There is a strong similarity between modeling parameters as random variables in random effects models and the Bayesian approach to statistical analysis. In mixed model analysis, however, some parameters are fixed, whereas in the Bayesian approach all parameters are treated as random variables. Many instances of Bayesian treatments of mixed or multilevel modeling are available in the literature; they are linked to Markov chain Monte Carlo methods and we discuss some of them in Chapter 6.

We let $\mathbf{y} = (y_1, \dots, y_n)^T$ denote the observed data vector. Conditional on the unobservable random effects vector, $\mathbf{u} = (u_1, \dots, u_q)^T$, we assume that \mathbf{y} arises from a GLM. The conditional mean $\mu_j = E(y_j | \mathbf{u})$ is related to the linear predictor $\eta_j = \mathbf{x}_j^T \boldsymbol{\beta} + \mathbf{z}_j^T \mathbf{u}$

by the link function $h(\mu_j) = \eta_j$ ($j = 1, \dots, n$), where β is a p -vector of fixed effects and x_j and z_j are, respectively, a p -vector and q -vector of explanatory variables associated with the fixed and random effects. This formulation encompasses the modeling of data involving multiple sources of random error, such as repeated measures within subjects and clustered data collected from some experimental units (Breslow and Clayton, 1993).

We let the distribution for \mathbf{u} be $p(\mathbf{u}; \mathbf{D})$, where \mathbf{D} denotes the vector of unknown parameters. The observed data in \mathbf{y} are conditionally independent with density functions of the form

$$f(y_j | \mathbf{u}; \beta, \kappa) = \exp[m_j \kappa^{-1}\{\theta_j y_j - b(\theta_j)\} + c(y_j; \kappa)], \quad (5.72)$$

where θ_j is the canonical parameter, κ is the dispersion parameter, m_j is the known prior weight, and b and c are scalar-valued functions. The conditional mean and canonical parameters are related through the equation $\mu_j = \partial b(\theta_j)/\partial \theta_j$. We let $\Psi = (\beta^T \kappa, \mathbf{D}^T)^T$.

The likelihood function for Ψ is given by

$$L(\Psi) = \int \prod_{j=1}^n f(y_j | \mathbf{u}; \beta, \kappa) p(\mathbf{u}; \mathbf{D}) d\mathbf{u}, \quad (5.73)$$

which cannot usually be evaluated in closed form due to an intractable integral whose dimension depends on the structure of the random effects. Within the EM framework, the random effects are considered as missing data. The complete-data vector \mathbf{x} is then given by $\mathbf{x} = (\mathbf{y}^T, \mathbf{u}^T)^T$ and the complete-data log likelihood is given by

$$\log L_c(\Psi) = \log L_{y,u}(\beta, \kappa) + \log p(\mathbf{u}; \mathbf{D}), \quad (5.74)$$

where

$$\log L_{y,u}(\beta, \kappa) = \sum_{j=1}^n \log f(y_j | \mathbf{u}; \beta, \kappa) \quad (5.75)$$

denotes the log conditional density of \mathbf{y} given \mathbf{u} . On the $(k+1)$ th iteration of the EM algorithm, the E-step involves the computation of the Q-function, $Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}}\{\log L_c(\Psi) | \mathbf{y}\}$, where the expectation is with respect to the conditional distribution of $\mathbf{u} | \mathbf{y}$ with current parameter value $\Psi^{(k)}$. As this conditional distribution involves the (marginal) likelihood function $L(\Psi)$ given in (5.73), an analytical evaluation of the Q-function for the model (5.72) will be impossible outside the normal theory mixed model (Booth and Hobert, 1999). The Monte Carlo EM algorithm can be adopted to tackle this problem, as to be discussed further in Section 6.3.4.

We now outline an approach based on the procedure of McGilchrist (1994) and McGilchrist and Yau (1995) to parameter estimation in GLMMs. It proceeds in two steps. We let $\theta = (\beta^T, \kappa, \mathbf{u}^T)^T$ and put

$$\Psi^{(0)} = (\beta^T, \kappa, \mathbf{D}^{(0)})^T,$$

where $\mathbf{D}^{(0)}$ is an initial estimate for \mathbf{D} . On the first step, an estimate $\theta^{(1)}$ for θ is obtained as a solution of the equation

$$\partial \log L_c(\Psi^{(0)}) / \partial \theta = \mathbf{0}. \quad (5.76)$$

Then the estimate $\mathbf{D}^{(0)}$ of \mathbf{D} is updated by its REML estimate, working with the unobservable vector \mathbf{u} of random effects replaced by its current estimate $\mathbf{u}^{(1)}$. This procedure is continued until convergence of the estimates is obtained. This approach has been developed further by Yau, Lee, and Ng (2003), Lee, Wang, Scott, Yau, and McLachlan (2006), Ng, McLachlan, Yau, and Lee (2004), and Xiang, Lee, Yau, and McLachlan (2006, 2007) for mixture modeling and estimation in a variety of situations. They use the EM algorithm to solve the equation (5.76) on each iteration to update the estimate of θ .

5.11 EXAMPLE 5.8: FACTOR ANALYSIS

Factor analysis is commonly used for explaining data, in particular, correlations between variables in multivariate observations. It can be used also for dimensionality reduction.

We let $\mathbf{W}_1, \dots, \mathbf{W}_n$ denote a random sample of size n on a p -dimensional random vector. In a typical factor analysis model, each observation \mathbf{W}_j is modeled as

$$\mathbf{W}_j = \boldsymbol{\mu} + \mathbf{B}\mathbf{a}_j + \mathbf{e}_j \quad (j = 1, \dots, n), \quad (5.77)$$

where \mathbf{a}_j is a q -dimensional ($q < p$) vector of latent or unobservable variables called factors and \mathbf{B} is a $p \times q$ matrix of factor loadings (parameters). It is assumed that

$$(\mathbf{W}_1^T, \mathbf{a}_1^T)^T, \dots, (\mathbf{W}_n^T, \mathbf{a}_n^T)^T$$

are i.i.d. The \mathbf{a}_j are assumed to be i.i.d. as $N(\mathbf{0}, \mathbf{I}_q)$, independently of the errors \mathbf{e}_j , which are assumed to be i.i.d. as $N(\mathbf{0}, \mathbf{D})$, where \mathbf{D} is a diagonal matrix,

$$\mathbf{D} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2),$$

and where \mathbf{I}_q denotes the $q \times q$ identity matrix. The σ_i^2 are called the uniquenesses. Thus, conditional on the \mathbf{a}_j , the \mathbf{W}_j are independently distributed as $N(\boldsymbol{\mu} + \mathbf{B}\mathbf{a}_j, \mathbf{D})$. Unconditionally, the \mathbf{W}_j are i.i.d. according to a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^T + \mathbf{D}. \quad (5.78)$$

Hence the MLE of $\boldsymbol{\mu}$ is the sample mean vector $\bar{\mathbf{w}}$.

Under the model (5.77), the variables in \mathbf{W}_j are conditionally independent given \mathbf{a}_j . Thus the factors in \mathbf{a}_j are intended to explain the correlations between the variables in \mathbf{W}_j , while the error terms \mathbf{e}_j represent the unexplained noise unique to a particular \mathbf{w}_j ($j = 1, \dots, n$). Note that in the case of $q > 1$, there is an infinity of choices for \mathbf{B} , since this model is still satisfied if we replace \mathbf{a}_j by $\mathbf{C}\mathbf{a}_j$ and \mathbf{B} by $\mathbf{B}\mathbf{C}^T$, where \mathbf{C} is any orthogonal matrix of order q . As $\frac{1}{2}q(q - 1)$ constraints are needed for \mathbf{B} to be uniquely defined, the number of free parameters is

$$pq + p - \frac{1}{2}q(q - 1); \quad (5.79)$$

see Lawley and Maxwell (1971, Chapter 1).

5.11.1 EM Algorithm for Factor Analysis

The factor analysis model (5.77) can be fitted by maximum likelihood, although the solution has to be computed iteratively as no closed-form expressions exist for the MLEs of \mathbf{B} and \mathbf{D} . They can be computed iteratively via the EM algorithm as considered in Dempster et al. (1977); see Rubin and Thayer (1982). We let Ψ be the vector of unknown parameters, containing the elements of $\boldsymbol{\mu}$, \mathbf{B} , and the diagonal elements of \mathbf{D} .

In order to apply the EM algorithm and its variants to this problem, we formulate

$$\mathbf{x} = (\mathbf{w}^T, \mathbf{a}_1^T, \dots, \mathbf{a}_n^T)^T$$

as the complete-data vector, where where $\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T$ is the observed data vector. Since $\hat{\boldsymbol{\mu}} = \bar{\mathbf{w}}$, we set $\boldsymbol{\mu} = \bar{\mathbf{w}}$ at the outset, which is equivalent to taking $\boldsymbol{\mu}^{(k)} = \bar{\mathbf{w}}$ on all iterations k in the subsequent application of the EM algorithm.

The complete-data log likelihood is, but for an additive constant,

$$\log L_c(\Psi) = -\frac{1}{2}n \log |\mathbf{D}| - \frac{1}{2} \sum_{j=1}^n \{(\mathbf{w}_j - \bar{\mathbf{w}} - \mathbf{B}\mathbf{a}_j)^T \mathbf{D}^{-1} (\mathbf{w}_j - \bar{\mathbf{w}} - \mathbf{B}\mathbf{a}_j) + \mathbf{a}_j^T \mathbf{a}_j\}.$$

The complete-data density belongs to the exponential family, and the complete-data sufficient statistics are \mathbf{C}_{ww} , \mathbf{C}_{wa} , and \mathbf{C}_{aa} , where

$$\mathbf{C}_{ww} = \sum_{j=1}^n (\mathbf{w}_j - \bar{\mathbf{w}})(\mathbf{w}_j - \bar{\mathbf{w}})^T; \quad \mathbf{C}_{wa} = \sum_{j=1}^n (\mathbf{w}_j - \bar{\mathbf{w}})\mathbf{a}_j^T; \quad \mathbf{C}_{aa} = \sum_{j=1}^n \mathbf{a}_j \mathbf{a}_j^T.$$

To calculate the conditional expectations of these sufficient statistics given the observed data \mathbf{y} , we need to use the result that the random vector $(\mathbf{W}_j^T, \mathbf{a}_j^T)^T$ has a multivariate normal distribution with mean

$$\begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix} \quad (5.80)$$

and covariance matrix

$$\begin{pmatrix} \mathbf{B}\mathbf{B}^T + \mathbf{D} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{I}_q \end{pmatrix}. \quad (5.81)$$

It thus follows that the conditional distribution of \mathbf{a}_j given \mathbf{w}_j is given by

$$\mathbf{a}_j | \mathbf{w}_j \sim N(\boldsymbol{\gamma}^T (\mathbf{w}_j - \boldsymbol{\mu}), \boldsymbol{\Omega}) \quad (5.82)$$

for $j = 1, \dots, n$, where

$$\boldsymbol{\gamma} = (\mathbf{B}\mathbf{B}^T + \mathbf{D})^{-1} \mathbf{B}. \quad (5.83)$$

and where

$$\boldsymbol{\Omega} = \mathbf{I}_q - \boldsymbol{\gamma}^T \mathbf{B}. \quad (5.84)$$

The EM algorithm is implemented as follows on the $(k+1)$ th iteration.

E-Step. Given the current fit $\Psi^{(k)}$ for Ψ , calculate as follows the conditional expectation of these sufficient statistics given the observed data \mathbf{y} :

$$E_{\Psi^{(k)}}(\mathbf{C}_{ww} | \mathbf{y}) = \mathbf{C}_{ww},$$

$$E_{\Psi^{(k)}}(\mathbf{C}_{wa} | \mathbf{y}) = \mathbf{C}_{ww} \boldsymbol{\gamma}^{(k)},$$

and

$$E_{\Psi^{(k)}}(\mathbf{C}_{aa} | \mathbf{y}) = \boldsymbol{\gamma}^{(k)T} \mathbf{C}_{ww} \boldsymbol{\gamma}^{(k)} + n\boldsymbol{\Omega}^{(k)},$$

where

$$\boldsymbol{\gamma}^{(k)} = \{\mathbf{B}^{(k)} \mathbf{B}^{(k)T} + \mathbf{D}^{(k)}\}^{-1} \mathbf{B}^{(k)}$$

and

$$\boldsymbol{\Omega}^{(k)} = \mathbf{I}_q - \boldsymbol{\gamma}^{(k)T} \mathbf{B}^{(k)}.$$

M-Step. Calculate

$$\begin{aligned} \mathbf{B}^{(k+1)} &= \mathbf{C}_{ww} \boldsymbol{\gamma}^{(k)} (\boldsymbol{\gamma}^{(k)T} \mathbf{C}_{ww} \boldsymbol{\gamma}^{(k)} + n\boldsymbol{\Omega}^{(k)})^{-1} \\ &= \mathbf{V} \boldsymbol{\gamma}^{(k)} (\boldsymbol{\gamma}^{(k)T} \mathbf{V} \boldsymbol{\gamma}^{(k)} + \boldsymbol{\Omega}^{(k)})^{-1}, \end{aligned} \quad (5.85)$$

where

$$\mathbf{V} = n^{-1} \mathbf{C}_{ww} \quad (5.86)$$

and

$$\begin{aligned} \mathbf{D}^{(k+1)} &= n^{-1} \operatorname{diag}\{\mathbf{C}_{ww} - \mathbf{C}_{ww} \boldsymbol{\gamma}^{(k)} (\boldsymbol{\gamma}^{(k)T} \mathbf{C}_{ww} \boldsymbol{\gamma}^{(k)} + n\boldsymbol{\Omega}^{(k)})^{-1} \boldsymbol{\gamma}^{(k)T} \mathbf{C}_{ww}\} \\ &= \operatorname{diag}\{\mathbf{V} - \mathbf{B}^{(k+1)} \mathbf{H}^{(k)} \mathbf{B}^{(k+1)T}\}, \end{aligned} \quad (5.87)$$

and where

$$\begin{aligned} \mathbf{H}^{(k)} &= (\boldsymbol{\gamma}^{(k)T} \mathbf{V} \boldsymbol{\gamma}^{(k)} + \boldsymbol{\Omega}^{(k)}) \\ &= n^{-1} E_{\boldsymbol{\Psi}^{(k)}}(\mathbf{C}_{aa} | \mathbf{y}) \\ &= n^{-1} \sum_{j=1}^n E_{\boldsymbol{\Psi}^{(k)}}(\mathbf{a}_j \mathbf{a}_j^T | \mathbf{w}_j). \end{aligned} \quad (5.88)$$

The inversion of the current value of the $p \times p$ matrix $(\mathbf{B}\mathbf{B}^T + \mathbf{D})$ on each iteration can be undertaken using the result that

$$(\mathbf{B}\mathbf{B}^T + \mathbf{D})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1} \mathbf{B} (\mathbf{I}_q + \mathbf{B}^T \mathbf{D}^{-1} \mathbf{B})^{-1} \mathbf{B}^T \mathbf{D}^{-1}, \quad (5.89)$$

where the right-hand side of (5.89) involves only the inverses of $q \times q$ matrices, since \mathbf{D} is a diagonal matrix. The determinant of $(\mathbf{B}\mathbf{B}^T + \mathbf{D})$ can then be calculated as

$$|\mathbf{B}\mathbf{B}^T + \mathbf{D}| = |\mathbf{D}| / |\mathbf{I}_q - \mathbf{B}^T(\mathbf{B}\mathbf{B}^T + \mathbf{D})^{-1}\mathbf{B}|.$$

Direct differentiation of the log likelihood function shows that the ML estimate of the diagonal matrix \mathbf{D} satisfies

$$\hat{\mathbf{D}} = \operatorname{diag}(\hat{\mathbf{V}} - \hat{\mathbf{B}}\hat{\mathbf{B}}^T). \quad (5.90)$$

As remarked by Lawley and Maxwell (1971, Page 30) in the context of direct computation of the ML estimate for a single-component factor analysis model, the equation (5.90) looks temptingly simple to use to solve for $\hat{\mathbf{D}}$, but was not recommended due to convergence problems.

On comparing (5.90) with (5.87), it can be seen that with the calculation of the ML estimate of \mathbf{D} directly from the (incomplete-data) log likelihood function, the unconditional expectation of $\mathbf{a}_j \mathbf{a}_j^T$, which is the identity matrix, is used in place of the conditional expectation in (5.87) on the E-step of the EM algorithm. Unlike the direct approach of calculating the ML estimate, the EM algorithm and its variants have good convergence properties in that they ensure the likelihood is not decreased after each iteration regardless of the choice of starting point; see McLachlan, Peel, and Bean (2003) for further discussion.

Although the EM algorithm is numerically stable, unlike the Newton-Raphson type algorithms for factor analysis, it has two unpleasant features. They are, as mentioned in Section 3.9.3, the possibility of multiple solutions (which are not necessarily rotations of each other) and slow convergence rate. The slowness of convergence could be attributed to the typically large fraction of missing data. Rubin and Thayer (1982, 1983) illustrate this EM algorithm with a data set from Lawley and Maxwell (1963), using some computations of Jöreskog (1969). They suggest that with the EM algorithm multiple local maxima can be reached from different starting points and question the utility of

second-derivative-based classical methods. Bentler and Tanaka (1983), however, question the multiplicity of solutions obtained and attribute it to the slowness of convergence. Horng (1987), as pointed out in Section 3.9.3, proves sublinear convergence of the EM algorithm in certain situations. Duan and Simonato (1993) investigate the issue of severity of the multiplicity of solutions with an examination of eight data sets and conclude that multiplicity is indeed a common phenomenon in factor analysis, supportive of the Bentler-Tanaka conclusion. They also show examples of different solutions obtained under different convergence criteria.

5.11.2 ECME Algorithm for Factor Analysis

Liu and Rubin (1994) develop the ECME algorithm for this problem. Here the partition of Ψ is $(\Psi_1^T, \Psi_2^T)^T$, where Ψ_1 contains the elements of B and $\Psi_2 = (\sigma_1^2, \dots, \sigma_p^2)^T$. This choice is motivated by the simplicity of numerical maximization of the actual log likelihood over the p -dimensional D , since it is the log likelihood for a normal distribution with restrictions on the covariance matrix compared to that of maximization over the matrix-valued B or over B and D . In this version of the ECME algorithm, the E-step is as in the EM algorithm above. There are two CM-steps.

CM-Step 1. Calculate $B^{(k+1)}$ as above.

CM-Step 2. Maximize the constrained actual log likelihood given $B^{(k+1)}$ by an algorithm such as Newton-Raphson.

5.11.3 Numerical Example

In the example from Lawley and Maxwell (1963), the lower half of the symmetric matrix C_{yy} as given by Rubin and Thayer (1982) is as follows. There are $p = 9$ variables.

$$C_{yy} = \begin{bmatrix} 1.0 & & & & & & & & \\ 0.554 & 1.0 & & & & & & & \\ 0.227 & 0.296 & 1.0 & & & & & & \\ 0.189 & 0.219 & 0.769 & 1.0 & & & & & \\ 0.461 & 0.479 & 0.237 & 0.212 & 1.0 & & & & \\ 0.479 & 0.530 & 0.243 & 0.226 & 0.520 & 1.0 & & & \\ 0.243 & 0.425 & 0.304 & 0.291 & 0.514 & 0.473 & 1.0 & & \\ 0.280 & 0.311 & 0.718 & 0.681 & 0.313 & 0.348 & 0.374 & 1.0 & 0.672 \\ 0.241 & 0.311 & 0.730 & 0.661 & 0.245 & 0.290 & 0.306 & 0.672 & 1.0 \end{bmatrix}$$

The model used is with $q = 4$. Liu and Rubin (1994) use the same starting values as in Rubin and Thayer (1982, 1983) and apply the above ECME algorithm. Results of the log likelihood and estimates of σ_3 over iterations are given in Figure 5.4 (EM in solid lines and ECME in dashed lines), from which it appears that the ECME algorithm converges much faster than the EM algorithm. In terms of CPU time, the ECME algorithm was only 25 percent faster than the EM algorithm.

5.11.4 EM Algorithm in Principal Component Analysis

The EM algorithm and its variants have also been used in principal component analysis, notably in the work of Bishop (1999), Tipping and Bishop (1999a, 1999b), Roweis (1997),

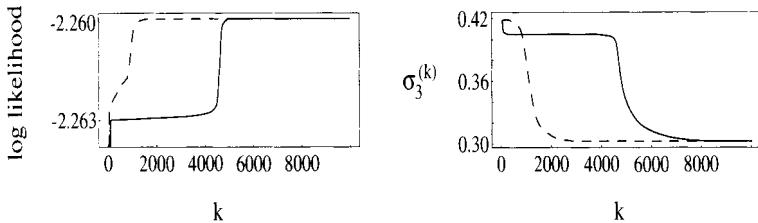


Figure 5.4 Factor Analysis: Convergence of EM (solid line) and ECME (dashed line) for log likelihood and $\sigma_3^{(k)}$. From Liu and Rubin, 1994.

and Schneider (2001). These algorithms are particularly useful when some observations are missing. The approaches followed in these papers are described in various sections of Jolliffe (2002).

In the context of principal component analysis and factor analysis, Tipping and Bishop (1999a) formulate a model for p -dimensional centered observations x_1, \dots, x_n as independent p -dimensional normal distributions with null-vector means and covariance matrix of the form $\mathbf{B}\mathbf{B}^T + \sigma^2 \mathbf{I}_p$, where \mathbf{B} is a $p \times q$ matrix. This is a particular case of the standard factor analysis model. They establish a connection between principal component analysis and factor analysis by showing that the MLE of \mathbf{B} is the matrix of the first q principal components obtained from the observations. The MLE of σ^2 is also related to the principal components—the mean of the smallest $(p - q)$ eigenvalues of the sample covariance matrix. The MLEs are derived using the EM algorithm considering the principal components as missing values. The EM algorithm may not really be necessary in this case, but comes in handy when there are missing observations in the actual sense or when a suitable model is a mixture of distributions. Bishop (1999) formulates a Bayesian version of the Tipping-Bishop model and the prior distribution on \mathbf{B} helps to decide on a suitable choice of q , the number of principal components. Roweis (1997) formulates a more general principal components analysis model where the error matrix $\sigma^2 \mathbf{I}_p$ is replaced by an arbitrary covariance matrix. He also uses the EM algorithm for MLE. The Tipping-Bishop special case arises when the error matrix is indeed of the form $\sigma^2 \mathbf{I}_p$ with $\sigma^2 \rightarrow 0$. Tipping and Bishop (1999b) consider a mixture model with several \mathbf{B}_g 's, one for each group and solves the estimation problem by a two-stage EM algorithm. They again relate the MLEs of \mathbf{B}_g 's to the first q principal components of the corresponding groups. Bishop (1999) tackles a Bayesian model for this mixture model as well.

By combining their version of the EM algorithm with Little and Rubin's (2002) (our Section 2.2) method of estimating multivariate normal parameters in the presence of missing values, they derive an algorithm for MLE of their model parameters when there are missing observations. An interesting aspect of their algorithm is that covariance matrices are not estimated as intermediate steps and the principal components are estimated directly.

By adapting an approach of adding a diagonal matrix to a current estimate of a covariance matrix like in ridge regression in an EM-type algorithm, Schneider (2001) describes a *regularized EM algorithm* for estimating covariance matrices. This approach is particularly useful when the number of variables exceeds the number of observations.

5.12 EFFICIENT DATA AUGMENTATION

5.12.1 Motivation

Meng and van Dyk (1997) consider speeding up the convergence of the EM algorithm through the choice of the complete-data, or effectively, the choice of the missing data in the specification of the complete-data problem in the EM framework. Their idea is to search for an efficient way of augmenting the observed data, where by efficient they mean less augmentation of the observed data while maintaining the stability and simplicity of the EM algorithm.

As shown in Section 3.9, the rate of convergence (equivalently, the speed of convergence) of the EM algorithm depends on the proportion of missing information in the prescribed EM framework. The smaller this proportion, the greater the speed of convergence. Hence by augmenting the observed data less, the speed of convergence of the resulting EM algorithm will be greater. However, a disadvantage of less augmentation of the observed data is that the resulting E- or M-steps, or both, may be made appreciably more difficult to implement. But if the E- and M-steps are equally (or only slightly less) simple to implement and the gain in speed is relatively substantial, then there is no reason not to use the faster EM algorithm. To this end, Meng and van Dyk (1997) introduce a working parameter in their specification of the complete data, which thus indexes a class of EM algorithms corresponding to the different values of the working parameter. The aim is to select a value of the working parameter that increases the speed of convergence (that is, provides less data augmentation), without appreciably affecting the stability and simplicity of the resulting EM algorithm.

5.12.2 Maximum Likelihood Estimation of t -Distribution

We shall now consider one of the examples that Meng and van Dyk (1997) use to illustrate their approach. It concerns the problem of ML estimation of the parameters of the multivariate t -distribution with known degrees of freedom ν , which was initially introduced as Example 2.6 in Section 2.6 and considered further in the previous sections of this chapter in the case of ν being also unknown.

Suppose as in Example 2.6 that $\mathbf{w}_1, \dots, \mathbf{w}_n$ denote an observed random sample from the multivariate t -distribution defined by (2.38). Then from (2.36), we can write \mathbf{W}_j as

$$\mathbf{W}_j = \boldsymbol{\mu} + \mathbf{C}_j/U_j^{1/2} \quad (j = 1, \dots, n), \quad (5.91)$$

where \mathbf{C}_j is distributed $N(\mathbf{0}, \Sigma)$ independently of U_j , which is distributed as

$$U_j \sim \text{gamma}(\frac{1}{2}\nu, \frac{1}{2}\nu). \quad (5.92)$$

From (5.91), \mathbf{W}_j can be expressed further as

$$\begin{aligned} \mathbf{W}_j &= \boldsymbol{\mu} + |\Sigma|^{-\frac{\alpha}{2}} \mathbf{C}_j / \{|\Sigma|^{-\alpha} U_j\}^{1/2} \\ &= \boldsymbol{\mu} + |\Sigma|^{-\frac{\alpha}{2}} \mathbf{C}_j / \{U_j(a)^{1/2}\}, \end{aligned} \quad (5.93)$$

where

$$U_j(a) = |\Sigma|^{-\alpha} U_j \quad (5.94)$$

and $U_j(0) = U_j$, as defined by (5.92).

Here a is the working parameter used by Meng and van Dyk (1997) to scale the missing data variable U_j ($j = 1, \dots, n$) in the complete-data formulation of the problem. The complete-data vector is defined as

$$\mathbf{x}(a) = (\mathbf{y}^T, \mathbf{z}^T(a))^T, \quad (5.95)$$

where

$$\mathbf{z}(a) = (u_1(a), \dots, u_n(a))^T$$

is the missing-data vector.

For a given value of a , the EM algorithm can be implemented with the complete data specified by (5.95). The case $a = 0$ corresponds to the standard specification as considered previously in Example 2.6. Because the distribution of $U_j(a)$ for $a \neq 0$ depends on Σ , the M -step of the EM algorithm, and hence its speed of convergence, is affected by the choice of a .

From (3.69) and (3.73), the global speed s of convergence of the EM algorithm in a neighborhood of a limit point Ψ^* is given by the smallest eigenvalue of

$$\mathcal{I}_c^{-1}(\Psi^*; \mathbf{y}) \mathbf{I}(\Psi^*; \mathbf{y}). \quad (5.96)$$

Here with the specification of the complete-data indexed by the working parameter a , s will be some function of a , $s(a)$. Meng and van Dyk (1997) suggest using the value of a , a_{opt} , that maximizes the speed $s(a)$; that is, a_{opt} is the value of a that maximizes the smallest eigenvalue of (5.96), in which the term $\mathbf{I}(\Psi^*; \mathbf{y})$ does not depend on a . They showed that a_{opt} is given by

$$a_{\text{opt}} = 1/(\nu + p), \quad (5.97)$$

which remarkably does not depend on the observed data \mathbf{y} .

We now consider the implementation of the EM algorithm for a given (arbitrary) value of the working parameter a . The implementation of the E-step is straightforward since the complete-data log likelihood is a linear function of $U_j(a)$, which is distributed as a (constant) multiple of U_j . Hence on the E-step at the $(k+1)$ th iteration, we have from (5.94) that

$$E_{\Psi^{(k)}}\{U_j(a) \mid \mathbf{y}\} = u_j^{(k)}(a),$$

where

$$u_j^{(k)}(a) = |\Sigma|^{-a} u_j^{(k)}, \quad (5.98)$$

and $u_j^{(k)}$ is given by (5.46).

The M-step is not straightforward for nonzero a , but rather fortuitously, its implementation simplifies in the case of $a = a_{\text{opt}}$. The only modification required to the current estimates of μ and Σ in the standard situation ($a = 0$) is to replace the divisor n by

$$\sum_{j=1}^n u_j^{(k)} \quad (5.99)$$

in the expression (2.49) for $\Sigma^{(k+1)}$.

This replacement does not affect the limit, as

$$\sum_{j=1}^n u_j^{(k)} \rightarrow n, \quad k \rightarrow \infty. \quad (5.100)$$

This was proved by Kent, Tyler, and Vardi (1994), who used it to modify one of their EM algorithms for fitting the t -distribution. They constructed an EM algorithm via a “curious likelihood identity” originally proposed in Kent and Tyler (1991) for transforming a p -dimensional location-scale t -likelihood into a $(p+1)$ -dimensional scale-only t -likelihood. They reported that this algorithm with the modification (5.99), converges faster than the EM algorithm for the usual specification of the complete data. Meng and van Dyk (1997) note that the EM algorithm with $a = a_{\text{opt}}$ is identical to the modified EM algorithm of Kent et al. (1994), hence explaining its faster convergence; see also Arslan, Constable, and Kent (1995).

We now outline the details of the M-step as outlined above for a given value of the working parameter a . The complete-data likelihood function $L_c(\Psi)$ formed from the complete-data vector $\mathbf{x}(a)$ can be factored into the product of the conditional density of $\mathbf{Y} = (\mathbf{W}_1, \dots, \mathbf{W}_n)^T$ given $\mathbf{z}(a)$ and the marginal density of $\mathbf{Z}(a)$. Unlike the case of $a = 0$ in Example 2.6, the latter density for nonzero a depends on Σ , being a function of $|\Sigma|^a$. As noted above, the conditional expectation of $\log L_c(\Psi)$ given the observed data \mathbf{y} , is effected by replacing $u_j(a)$ by its current conditional expectation, $u_j^{(k)}(a)$. It follows on ignoring terms not depending on Ψ that the Q -function is given by

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) &= \frac{1}{2} n \log |\Sigma| \{a(p + \nu - 1)\} \\ &\quad - \frac{1}{2} |\Sigma|^a \sum_{j=1}^n u_j^{(k)}(a) \{\nu + (\mathbf{w}_j - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{w}_j - \boldsymbol{\mu})\}. \end{aligned} \tag{5.101}$$

The term

$$\sum_{j=1}^n u_j^{(k)}(a) (\mathbf{w}_j - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{w}_j - \boldsymbol{\mu})$$

on the right-hand side of (5.101) can be written as

$$\sum_{j=1}^n u_j^{(k)}(a) \{(\bar{\mathbf{w}}_u^{(k)} - \boldsymbol{\mu})^T \Sigma^{-1} (\bar{\mathbf{w}}_u^{(k)} - \boldsymbol{\mu}) + \text{tr}(\Sigma^{-1} S_u^{(k)})\}, \tag{5.102}$$

where

$$\begin{aligned} \bar{\mathbf{w}}^{(k)} &= \sum_{j=1}^n u_j^{(k)}(a) \mathbf{w}_j / \sum_{j=1}^n u_j^{(k)}(a), \\ &= \sum_{j=1}^n u_j^{(k)} \mathbf{w}_j / \sum_{j=1}^n u_j^{(k)}, \end{aligned}$$

and

$$\begin{aligned} S_u^{(k)} &= \sum_{j=1}^n u_j^{(k)}(a) (\mathbf{w}_j - \bar{\mathbf{w}}_u^{(k)}) (\mathbf{w}_j - \bar{\mathbf{w}}_u^{(k)})^T / \sum_{j=1}^n u_j^{(k)}(a) \\ &= \sum_{j=1}^n u_j^{(k)} (\mathbf{w}_j - \bar{\mathbf{w}}_u^{(k)}) (\mathbf{w}_j - \bar{\mathbf{w}}_u^{(k)})^T / \sum_{j=1}^n u_j^{(k)}. \end{aligned} \tag{5.103}$$

It follows from (5.101) and (5.102) that

$$\boldsymbol{\mu}^{(k+1)} = \bar{\mathbf{w}}_u^{(k)}.$$

Concerning the calculation of $\Sigma^{(k+1)}$, Meng and van Dyk (1997) show that on differentiation of (5.101) with respect to the elements of Σ^{-1} and putting $\boldsymbol{\mu} = \boldsymbol{\mu}^{(k+1)} = \bar{\mathbf{w}}_u^{(k)}$, it satisfies the equation

$$\frac{\{a(p + \nu) - 1\}}{|\Sigma|^a \bar{u}^{(k)}(a)} \Sigma + \mathbf{S}_u^{(k)} = a\{\nu + \text{tr}(\Sigma^{-1} \mathbf{S}_u^{(k)})\} \Sigma. \quad (5.104)$$

where

$$\bar{u}^{(k)}(a) = \sum_{j=1}^n u_j^{(k)}(a)/n.$$

As remarked by Meng and van Dyk (1997), solving (5.104) for arbitrary a is quite difficult, but there are two values of a that make (5.104) trivial to solve. One is $a = 0$, while the other is $a = a_{\text{opt}}$, for which the first term on the left-hand side of (5.104) is zero, which directly shows that $\Sigma^{(k+1)}$ is proportional to $\mathbf{S}_u^{(k)}$. It then follows that

$$\Sigma^{(k+1)} = \mathbf{S}_u^{(k)}.$$

It will be seen in Section 5.15 that the choice of $a = a_{\text{opt}}$ corresponds to an application of the parameter-expanded EM (PX-EM) algorithm of Liu et al. (1998).

We are presently only considering the case where the degrees of freedom ν is known. The extension to the general case of unknown ν is straightforward. For, since the presence of the working parameter a does not affect the computation of the parameter ν , it is calculated as in the case of $a = 0$ in Section 5.8.

As the theoretical measure $s(a)$ only measures the speed of convergence near convergence, Meng and van Dyk (1997) perform some simulations to further examine the actual gains in computational time by using $a = a_{\text{opt}}$ over $a = 0$ in fitting the t -distribution with known degrees of freedom ν . Random samples of size $n = 100$ were generated from each of three univariate distributions: (1) the standard normal; (2) a t -distribution with $\mu = 0$, $\sigma^2 = 1$, and $\nu = 1$ (that is, the Cauchy distribution); and (3) a mixture of a standard normal and an exponential with mean 3 in proportions 2/3 and 1/3. The simulations from distributions (1) and (3) were intended to reflect the fact that, in reality, there is no guarantee that the data are from a t -distribution, nor even a symmetric model. For each configuration, there were 1,000 simulation trials, on each of which the EM algorithm for $a = 0$ and a_{opt} was started with $\boldsymbol{\mu}$ and σ^2 equal to the sample mean and variance, respectively. The number of iterations N_0 and N_{opt} taken by the EM algorithm in achieving

$$\|\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^{(k)}\|^2 / \|\boldsymbol{\Psi}^{(k)}\|^2 \leq 10^{-10}$$

for $a = 0$ and a_{opt} , respectively, was recorded. A comparison of the number of iterations taken by two algorithms can be misleading, but in this case the EM algorithm for both values of a require the same amount of computation per iteration. It was found that on all 6000 simulation trials the EM algorithm for $a = a_{\text{opt}}$ was faster than for $a = 0$. Generally, the improvement was quite significant. On 5997 trials, the improvement was greater than 10 percent, and often reached as high as 50 percent when the Cauchy distribution was being fitted. Meng and van Dyk (1997) note that the EM algorithm in the standard situation of

$a = 0$ tends to be slower for smaller values of ν in the t -distribution, and thus the observed improvement is greatest where it is most useful.

Meng and van Dyk (1997) also perform a second simulation experiment to assess the gains in higher dimensions. In this experiment, 1000 random samples of size $n = 100$ were generated from a $p = 10$ dimensional Cauchy distribution. The EM algorithm for the standard situation of $a = 0$ was found to be at least 6.5 times slower on every trial and was usually between 8 to 10 times as slow. As explained by Meng and van Dyk (1997), the difference between the two versions of the EM algorithm corresponding to $a = 0$ and $a = a_{\text{opt}}$, stems from the fact that the Q -function, $Q(\Psi; \Psi^{(k)})$ in the latter case is flatter because of less data augmentation, and so provides a better approximation to the incomplete-data log likelihood, $\log L(\Psi)$.

5.12.3 Variance Components Model

Meng and van Dyk (1997) also consider the efficient choice of the complete data in the variance components model (5.59), to which the ECME was applied in Section 5.9.1. Analogous to the rescaling of the missing data through the use of a working parameter a in the t -model considered above, Meng and van Dyk (1997) consider rescaling the vector \mathbf{b} of random effects by \mathbf{D}^{-a} , where \mathbf{D} is the covariance matrix of \mathbf{b} and a is an arbitrary constant. The missing data vector is then declared to be

$$\mathbf{z}(a) = |\mathbf{D}|^{-a} \mathbf{b}.$$

However, as the resulting EM algorithm would be very difficult for arbitrary \mathbf{D} , Meng and van Dyk (1997) first diagonalize \mathbf{D} . Let \mathbf{C} be a lower triangular matrix such that

$$\mathbf{C}\mathbf{D}\mathbf{C}^T = \text{diag}(d_1, \dots, d_n). \quad (5.105)$$

Then $\mathbf{b}^* = \mathbf{C}\mathbf{b}$ has the diagonal covariance matrix given by (5.105). Meng and van Dyk (1997) then specify the missing data vector as

$$\mathbf{z}(a) = (b_1^*/d_1^{a_1/2}, \dots, b_n^*/d_n^{a_n/2})^T,$$

where the working parameter

$$\mathbf{a} = (a_1, \dots, a_n)^T$$

is now a vector so as to allow each transformed random effect to be scaled by its own standard deviation. Although, in principle, the EM algorithm can be implemented for any given value of \mathbf{a} , Meng and van Dyk (1997) restrict each a_i to being zero or one in order to keep the resulting EM algorithm simple to implement. They report empirical and theoretical results demonstrating the gains that were achievable with this approach.

5.13 ALTERNATING ECM ALGORITHM

Meng and van Dyk (1997) propose an extension of the EM algorithm called the Alternating ECM (AECM) algorithm. It is really an extension of the ECM algorithm, where the specification of the complete-data is allowed to be different on each CM-step. That is, the complete-data vector need not be the same on each CM-step. We have seen that the ECME algorithm allows one to consider maximization of the actual log likelihood function on a CM-step rather than the Q -function, representing the current conditional expectation of the

complete-data log likelihood function. Hence it can be viewed as a special case of the AECM algorithm, where the complete data can be specified as the observed data \mathbf{y} on a CM-step.

There is also the Space-Alternating Generalized EM (SAGE) algorithm proposed by Fessler and Hero (1994). Their algorithm, which was developed without knowledge of the ECM or ECME algorithms, was motivated from a different angle to that of the ECM algorithm of Meng and Rubin (1993). They started with the CM algorithm (that is, there is no initial augmentation of the observed data) in the special case where each step corresponds to a partition of the parameter vector into subvectors. They then propose the EM algorithm be applied to implement any CM-step for which an explicit solution does not exist. The specification of the complete-data is therefore usually different for each CM-step. With the SAGE algorithm, only one EM iteration is performed on a given CM-step, which is in the same spirit of the ECM algorithm which performs only one iteration of the CM algorithm on each EM iteration.

Meng and van Dyk (1997) propose generalizing the SAGE and ECME algorithms by combining them into the one algorithm, called the AECM algorithm. It allows the augmentation of the observed data to vary over the CM-steps, which is a key ingredient of the SAGE algorithm. It also allows for CM-steps that go beyond a simple partition of the parameter vector into subvectors (as, for example, in Section 5.6), a key feature of the ECME algorithm.

Meng and van Dyk (1997) establish that monotone convergence of the sequence of likelihood values $L(\Psi^{(k)})$ is retained with the AECM algorithm and that under standard regularity conditions, the sequence of parameter iterates converges to a stationary point of $L(\Psi)$. They also provide more complete results for the ECME and SAGE algorithms.

Meng and van Dyk (1997) illustrate the AECM algorithm by applying it to the ML estimation of the parameters of the t -distribution with all parameters unknown. The application of the ECME algorithm to this problem was described in Section 5.8. They consider two versions of this algorithm for this problem. As with the aforementioned application of the ECME algorithm to this problem, both versions had two CM-steps as set out in Section 5.9, corresponding to the partition,

$$\Psi = (\Psi_1^T, \Psi_2)^T,$$

where Ψ_1 contains μ and the distinct elements of Σ , and with Ψ_2 a scalar equal to ν . In Version 1 of the AECM algorithm, the complete-data vector is specified on both CM-steps by (5.95) with $a = a_{\text{opt}}$ as given by (5.97). In Version 2, the complete-data vector is taken to be the observed-data vector \mathbf{y} on the second CM-step for the calculation of the current estimate of ν , as in the application of the ECME algorithm to this problem.

Meng and van Dyk (1997) performed some simulations to compare these two versions of the AECM algorithm with the multicycle ECM and ECME algorithms. Random samples of size $n = 100$ were generated from a $p = 10$ dimensional t -distribution with $\mu = \mathbf{0}$ and $\nu = 10$, and where Σ was selected at the outset of the simulation as a positive definite nondiagonal matrix. The same stopping criterion and the same starting values for μ and Σ as in their simulations described in Section 5.12.2 were used for the four algorithms. The starting value for ν was $\nu^{(0)} = 10$. Version 1 of the AECM algorithm was found to be eight to twelve times faster than either the multicycle ECM or the ECME algorithms. The cost per iteration is less for Version 1 and the multicycle ECM algorithm than for the ECME algorithm. Moreover, the Version 2 of the AECM algorithm was only slightly more efficient than Version 1 in terms of the number of iterations required, and less efficient in terms of actual computer time.

In these results the choice of the complete-data vector for the computation of ν on the second CM-step makes little difference in terms of the number of iterations required for convergence. Yet, as Meng and van Dyk (1997) point out, Liu and Rubin (1994, 1995) have shown that the ECME algorithm can be much more efficient than the multicycle ECM algorithm. However, there are two principal differences between their examples and the simulations of Meng and van Dyk (1997). The latter were for ten-dimensional data with no missing values, whereas the examples of Liu and Rubin (1994, 1995) were for bivariate data, which has a very high proportion of missing values.

Meng and van Dyk (1997) replicate the analyses in Liu and Rubin (1995) to investigate the relative merit of the four algorithms in the presence of missing values. These analyses were for the clinical-trial example of Liu and Rubin (1995), where the four-dimensional data were taken from Shih and Weisberg (1986), and the artificial bivariate data set of Murray (1977). As above, Version 1 of the AECM algorithm emerges clearly as their recommended choice.

In other illustrative work on the AECM algorithm, Meng and van Dyk (1997) apply it to the analysis of PET/SPECT data. The application of the EM algorithm to this problem has been considered in Section 2.5.

5.14 EXAMPLE 5.9: MIXTURES OF FACTOR ANALYZERS

We illustrate here the use of the AECM algorithm by using it to fit mixtures of factor analyzers. We firstly consider the case of normal component factor analyzers before proceeding to consider t -component factor analyzers.

As demonstrated in McLachlan and Peel (2000a, Chapter 8), factor analysis can be used for dimensionality reduction. However, a single-factor analysis model like a principal component analysis, provides only a global linear model for the representation of the data in a lower-dimensional subspace. Thus it has limited scope in revealing group structure in a data set.

A global nonlinear approach can be obtained by postulating a finite mixture of linear submodels for the distribution of the full observation vector \mathbf{W}_j given the (unobservable) factors \mathbf{a}_j . That is, we can provide a local dimensionality reduction method by assuming that the distribution of the observation \mathbf{W}_j can be modeled as

$$\mathbf{W}_j = \boldsymbol{\mu}_i + \mathbf{B}_i \mathbf{a}_{ij} + \mathbf{e}_{ij} \quad \text{with prob. } \pi_i \quad (i = 1, \dots, g) \quad (5.106)$$

for $j = 1, \dots, n$, where \mathbf{a}_{ij} is a q -dimensional ($q < p$) vector of latent or unobservable variables called factors and \mathbf{B} is a $p \times q$ matrix of factor loadings (parameters). The factor (vector) \mathbf{a}_{ij} is distributed $N_q(\mathbf{0}, \mathbf{I}_q)$, independently of \mathbf{e}_{ij} , which is distributed $N_p(\mathbf{0}, \mathbf{D}_i)$, where \mathbf{D}_i is a diagonal matrix ($i = 1, \dots, g$) and where \mathbf{I}_q denotes the $q \times q$ identity matrix.

Thus the mixture of factor analyzers model is given by

$$f(\mathbf{w}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \sum_{i=1}^g \pi_i \phi(\mathbf{w}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (5.107)$$

where the i th component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form

$$\boldsymbol{\Sigma}_i = \mathbf{a}_i \mathbf{a}_i^T + \mathbf{D}_i \quad (i = 1, \dots, g), \quad (5.108)$$

and $\phi(\mathbf{w}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the multivariate normal density function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The parameter vector $\boldsymbol{\Psi}$ now consists of the elements of the $\boldsymbol{\mu}_i$,

the \mathbf{B}_i , and the \mathbf{D}_i , along with the mixing proportions π_i ($i = 1, \dots, g - 1$), on putting $\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i$. In the above formulation (5.108), the factors \mathbf{a}_i are specific to the components, but the EM algorithm can be implemented with no extra difficulty without this specification to give the same results.

Unlike the principal component analysis model, the mixture of factor analyzers model (5.106) enjoys a powerful invariance property: changes in the scales of the feature variables in \mathbf{y}_j , appear only as scale changes in the appropriate rows of the matrix \mathbf{B}_i of factor loadings (in conjunction with scale changes in the elements of the vectors of means and errors.) This approach has been studied in a series of articles by McLachlan and Peel (2000a, 2000b), Peel and McLachlan (2000), and McLachlan et al. (2003).

5.14.1 Normal Component Factor Analyzers

In Example 5.8 in Section 5.11, we showed how the ECM algorithm can be applied to fit a single factor analyzer to an observed random sample,

$$\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T.$$

We now consider the fitting of the mixture of factor analyzers to the data in \mathbf{y} , using a multicycle AECM algorithm.

To apply the AECM algorithm to the fitting of the mixture of factor analyzers model, we partition the vector of unknown parameters Ψ as $(\Psi_1^T, \Psi_2^T)^T$, where Ψ_1 contains the mixing proportions π_i ($i = 1, \dots, g - 1$) and the elements of the component means μ_i ($i = 1, \dots, g$). The subvector Ψ_2 contains the elements of the \mathbf{B}_i and the \mathbf{D}_i ($i = 1, \dots, g$).

We let $\Psi^{(k)} = (\Psi_1^{(k)T}, \Psi_2^{(k)T})^T$ be the value of Ψ after the k th iteration of the AECM algorithm. For this application of the AECM algorithm, one iteration consists of two cycles, and there is one E-step and two CM-steps for each cycle. The two CM-steps correspond to the partition of Ψ into the two subvectors Ψ_1 and Ψ_2 .

For the first cycle of the AECM algorithm, we specify the missing data to be just the component-indicator vectors, $\mathbf{z}_1, \dots, \mathbf{z}_n$, where $z_{ij} = (\mathbf{z}_j)_i$ is one or zero, according to whether \mathbf{w}_j arose or did not arise from the i th component ($i = 1, \dots, g$; $j = 1, \dots, n$). In this conceptualization of the mixture model, it is valid to assume that the observation \mathbf{y}_j has arisen from one of the g components. For the second cycle for the updating of Ψ_2 , we specify the missing data to be the factors $\mathbf{a}_{i1}, \dots, \mathbf{a}_{in}$, as well as the component-indicator labels z_{ij} .

5.14.2 E-step

In order to carry out the E-step, we need to be able to compute the conditional expectation of the sufficient statistics. To carry out this step, we need to be able to calculate the conditional expectations,

$$E\{Z_{ij}\mathbf{w}_j\mathbf{a}_{ij}^T \mid \mathbf{w}_j\} \tag{5.109}$$

and

$$E\{Z_{ij}\mathbf{a}_{ij}\mathbf{a}_{ij}^T \mid \mathbf{w}_j\}. \tag{5.110}$$

It follows from (5.82) that the conditional distribution of \mathbf{a}_{ij} given \mathbf{w}_j and $z_{ij} = 1$ is given by

$$\mathbf{a}_j \mid \mathbf{w}_j, z_{ij} = 1 \sim N(\boldsymbol{\gamma}_i^T(\mathbf{w}_j - \boldsymbol{\mu}_i), \boldsymbol{\Omega}_i) \tag{5.111}$$

for $i = 1, \dots, g; j = 1, \dots, n$, where

$$\boldsymbol{\gamma}_i = (\mathbf{B}_i \mathbf{B}_i^T + \mathbf{D}_i)^{-1} \mathbf{B}_i. \quad (5.112)$$

and where

$$\boldsymbol{\Omega}_i = \mathbf{I}_q - \boldsymbol{\gamma}_i^T \mathbf{B}_i. \quad (5.113)$$

Using (5.111),

$$E\{Z_{ij} \mathbf{w}_j \mathbf{a}_{ij}^T \mid \mathbf{w}_j\} = \tau_i(\mathbf{w}_j; \boldsymbol{\Psi}) \mathbf{w}_j (\mathbf{w}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\gamma}_i \quad (5.114)$$

and

$$E\{\mathbf{a}_{ij} \mathbf{a}_{ij}^T \mid \mathbf{w}_j\} = \tau_i(\mathbf{w}_j; \boldsymbol{\Psi}) \{\boldsymbol{\gamma}_i^T (\mathbf{w}_j - \boldsymbol{\mu}_i) (\mathbf{w}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\gamma}_i + \boldsymbol{\Omega}_i\}, \quad (5.115)$$

where $\tau_i(\mathbf{w}_j; \boldsymbol{\Psi})$ is the i th component-posterior probability of \mathbf{w}_j defined by

$$\tau_{ij}(\mathbf{w}_j; \boldsymbol{\Psi}) = \frac{\pi_i \phi(\mathbf{w}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{h=1}^g \pi_h \phi(\mathbf{w}_j; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}. \quad (5.116)$$

5.14.3 CM-steps

The first conditional CM-step leads to $\pi_i^{(k)}$ and $\boldsymbol{\mu}_i^{(k)}$ being updated to

$$\pi_i^{(k+1)} = \sum_{j=1}^n \tau_i(\mathbf{w}_j; \boldsymbol{\Psi}^{(k)}) / n \quad (5.117)$$

and

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^n \tau_i(\mathbf{w}_j; \boldsymbol{\Psi}^{(k)}) \mathbf{w}_j / \sum_{j=1}^n \tau_i(\mathbf{w}_j; \boldsymbol{\Psi}^{(k)}) \quad (5.118)$$

for $i = 1, \dots, g$.

For the second cycle for the updating of $\boldsymbol{\Psi}_2$, we specify the missing data to be the factors $\mathbf{a}_{i1}, \dots, \mathbf{a}_{in}$, as well as the component-indicator vectors, $\mathbf{z}_1, \dots, \mathbf{z}_n$. On setting $\boldsymbol{\Psi}^{(k+1/2)}$ equal to $(\boldsymbol{\Psi}_1^{(k+1)^T}, \boldsymbol{\Psi}_2^{(k)^T})^T$, an E-step is performed to calculate $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k+1/2)})$, which is the conditional expectation of the complete-data log likelihood given the observed data, using $\boldsymbol{\Psi} = \boldsymbol{\Psi}^{(k+1/2)}$. The CM-step on this second cycle is implemented by the maximization of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k+1/2)})$ over $\boldsymbol{\Psi}$ with $\boldsymbol{\Psi}_1$ set equal to $\boldsymbol{\Psi}_1^{(k+1)}$. This yields the updated estimates $\mathbf{B}_i^{(k+1)}$ and $\mathbf{D}_i^{(k+1)}$. The former is given by

$$\mathbf{B}_i^{(k+1)} = \mathbf{V}_i^{(k+1/2)} \boldsymbol{\gamma}_i^{(k)} (\boldsymbol{\gamma}_i^{(k)^T} \mathbf{V}_i^{(k+1/2)} \boldsymbol{\gamma}_i^{(k)} + \boldsymbol{\Omega}_i^{(k)})^{-1}, \quad (5.119)$$

where

$$\mathbf{V}_i^{(k+1/2)} = \frac{\sum_{j=1}^n \tau_i(\mathbf{w}_j; \boldsymbol{\Psi}^{(k+1/2)}) (\mathbf{w}_j - \boldsymbol{\mu}_i^{(k+1)}) (\mathbf{w}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^n \tau_i(\mathbf{w}_j; \boldsymbol{\Psi}^{(k+1/2)})}, \quad (5.120)$$

$$\boldsymbol{\gamma}_i^{(k)} = (\mathbf{B}_i^{(k)} \mathbf{B}_i^{(k)^T} + \mathbf{D}_i^{(k)})^{-1} \mathbf{B}_i^{(k)}, \quad (5.121)$$

and

$$\Omega_i^{(k)} = \mathbf{I}_q - \boldsymbol{\gamma}_i^{(k)T} \mathbf{B}_i^{(k)} \quad (5.122)$$

for $i = 1, \dots, g$. The updated estimate $\mathbf{D}_i^{(k+1)}$ is given by

$$\mathbf{D}_i^{(k+1)} = \text{diag}\{\mathbf{V}_i^{(k+1/2)} - \mathbf{B}_i^{(k+1)} \mathbf{H}_i^{(k+1/2)} \mathbf{B}_i^{(k+1)T}\}, \quad (5.123)$$

where

$$\begin{aligned} \mathbf{H}_i^{(k+1/2)} &= \frac{\sum_{j=1}^n \tau_i(\mathbf{w}_j; \Psi^{(k+1/2)}) E_i^{(k+1/2)}(\mathbf{a}_j \mathbf{a}_j^T | \mathbf{w}_j)}{\sum_{j=1}^n \tau_i(\mathbf{w}_j; \Psi^{(k+1/2)})} \\ &= \boldsymbol{\gamma}_i^{(k)T} \mathbf{V}_i^{(k+1/2)} \boldsymbol{\gamma}_i^{(k)} + \Omega_i^{(k)} \end{aligned} \quad (5.124)$$

and $E_i^{(k+1/2)}$ denotes conditional expectation given membership of the i th component, using $\Psi^{(k+1/2)}$ for Ψ .

Some of the estimates of the elements of the diagonal matrix \mathbf{D}_i (the uniquenesses) will be close to zero if effectively not more than q observations are unequivocally assigned to the i th component of the mixture in terms of the fitted posterior probabilities of component membership. This will lead to spikes or near singularities in the likelihood. One way to avoid this is to impose the condition of a common value \mathbf{D} for the \mathbf{D}_i ,

$$\mathbf{D}_i = \mathbf{D} \quad (i = 1, \dots, g). \quad (5.125)$$

Another way is to impose constraints on the ratios of the diagonal elements of each \mathbf{D}_i ; see Hathaway (1985) and Ingrassia (2004). Alternatively, one can adopt a Bayesian approach as, for example, in Fokoué and Titterington (2002) and Svensén and Bishop (2005).

Under the mixture of probabilistic component analyzers (PCAs) model as proposed by Tipping and Bishop (1997), the i th component-covariance matrix Σ_i has the form (5.108) with each \mathbf{D}_i now having the isotropic structure

$$\mathbf{D}_i = \sigma_i^2 \mathbf{I}_p \quad (i = 1, \dots, g). \quad (5.126)$$

Under this isotropic restriction (2.7), $\mathbf{B}_i^{(k+1)}$ and $\sigma_i^{(k+1)^2}$ are given explicitly by an eigenvalue decomposition of the current value of \mathbf{V}_i without the need to introduce the latent factors \mathbf{u}_{ij} as “missing” data.

We can make use of the link of factor analysis with the probabilistic PCA algorithm to specify an initial starting value for Ψ ; see McLachlan et al. (2003).

5.14.4 *t*-Component Factor Analyzers

The mixture of factor analyzers model is sensitive to outliers since it adopts the multivariate normal family for the distributions of the errors and the latent factors. An obvious way to improve the robustness of this model for data which have longer tails than the normal or atypical observations is to consider using the multivariate *t*-family of elliptically symmetric distributions. It has an additional parameter called the degrees of freedom that controls the length of the tails of the distribution. Although the number of outliers needed for breakdown is almost the same as with the normal distribution, the outliers have to be much larger (Hennig, 2004).

We now formulate our mixture of t -analyzers model by replacing the multivariate normal distribution in (5.107) for the i th component-conditional distribution of \mathbf{W}_j by the multivariate t -distribution with mean vector $\boldsymbol{\mu}_i$, scale matrix $\boldsymbol{\Sigma}_i$, and ν_i degrees with the factor analytic restriction (5.108) on the component-scale matrices $\boldsymbol{\Sigma}_i$. Thus our postulated mixture model of t -factor analyzers assumes that $\mathbf{w}_1, \dots, \mathbf{w}_n$ is an observed random sample from the t -mixture density

$$f(\mathbf{w}_j; \boldsymbol{\Psi}) = \sum_{i=1}^g \pi_i f_p(\mathbf{w}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i), \quad (5.127)$$

where

$$\boldsymbol{\Sigma}_i = \mathbf{B}_i \mathbf{B}_i^T + \mathbf{D}_i \quad (i = 1, \dots, g) \quad (5.128)$$

and where the t -mixture density is defined by (2.38). The vector of unknown parameters $\boldsymbol{\Psi}$ consists of the degrees of freedom ν_i in addition to the mixing proportions π_i and the elements of the $\boldsymbol{\mu}_i$, \mathbf{B}_i , and the \mathbf{D}_i ($i = 1, \dots, g$). As in the mixture of normal factor analyzers model, \mathbf{B}_i is a $p \times q$ matrix and \mathbf{D}_i is a diagonal matrix. Zhao and Jiang (2006) considered this problem in the special case of spherical \mathbf{D}_i . McLachlan, Bean, and Ben-Tovim Jones (2007) considered the general case as presented here.

Corresponding to (5.106), we assume that

$$\mathbf{W}_j = \boldsymbol{\mu}_i + \mathbf{B}_i \mathbf{a}_{ij} + \mathbf{e}_{ij} \quad \text{with prob. } \pi_i \quad (i = 1, \dots, g) \quad (5.129)$$

for $j = 1, \dots, n$, where the joint distribution of the factor \mathbf{a}_{ij} and of the error \mathbf{e}_{ij} needs to be specified so that it is consistent with the t -mixture formulation (5.127) for the marginal distribution of \mathbf{W}_j .

From (5.80) and (5.81), we have for the usual factor analysis model that conditional on membership of the i th component of the mixture the joint distribution of \mathbf{W}_j and its associated factor (vector) \mathbf{a}_{ij} is multivariate normal,

$$\begin{pmatrix} \mathbf{W}_j \\ \mathbf{a}_{ij} \end{pmatrix} \mid z_{ij} = 1 \sim N_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i) \quad (i = 1, \dots, g), \quad (5.130)$$

where the mean $\boldsymbol{\mu}_i^*$ and the covariance matrix $\boldsymbol{\xi}_i$ are defined by

$$\boldsymbol{\mu}_i^* = (\boldsymbol{\mu}_i^T, \mathbf{0}^T)^T \quad (5.131)$$

and

$$\boldsymbol{\xi}_i = \begin{pmatrix} \mathbf{B}_i \mathbf{B}_i^T + \mathbf{D}_i & \mathbf{B}_i \\ \mathbf{B}_i^T & \mathbf{I}_q \end{pmatrix}. \quad (5.132)$$

We now replace the normal distribution by the t -distribution in (5.130) to postulate that

$$\begin{pmatrix} \mathbf{W}_j \\ \mathbf{a}_{ij} \end{pmatrix} \mid z_{ij} = 1 \sim t_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i, \nu_i) \quad (i = 1, \dots, g). \quad (5.133)$$

This specification of the joint distribution of \mathbf{W}_j and its associated factors in (5.129) will imply the t -mixture model (5.127) for the marginal distribution of \mathbf{W}_j with the restriction (5.128) on its component-scale matrices $\boldsymbol{\Sigma}_i$. Using the characterization of the t -distribution discussed in Section 2.6.1, it follows that we can express (5.130) alternatively as

$$\begin{pmatrix} \mathbf{W}_j \\ \mathbf{a}_{ij} \end{pmatrix} \mid u_j, z_{ij} = 1 \sim N_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i/u_j), \quad (5.134)$$

where u_j is a value of the weight variable U_j taken to have the gamma ($\frac{1}{2}\nu_i, \frac{1}{2}\nu_i$). It can be established from (5.134) that

$$\mathbf{a}_{ij} | u_j, z_{ij} = 1 \sim N_q(\mathbf{0}, \mathbf{I}_q/u_j) \quad (5.135)$$

and

$$\mathbf{e}_{ij} | u_j, z_{ij} = 1 \sim N_p(\mathbf{0}, \mathbf{D}_i/u_j), \quad (5.136)$$

and hence that

$$\mathbf{a}_{ij} | z_{ij} = 1 \sim t_q(\mathbf{0}, \mathbf{I}_q, \nu_i) \quad (5.137)$$

and

$$\mathbf{e}_{ij} | z_{ij} = 1 \sim t_p(\mathbf{0}, \mathbf{D}_i, \nu_i). \quad (5.138)$$

Thus with this formulation, the error terms \mathbf{e}_{ij} and the factors \mathbf{a}_{ij} are distributed according to the t -distribution with the same degrees of freedom. However, the factors and error terms are no longer independently distributed as in the normal-based model for factor analysis, but they are uncorrelated. To see this, we have from (5.134) that conditional on w_j , \mathbf{a}_{ij} and \mathbf{e}_{ij} are uncorrelated, and hence, unconditionally uncorrelated.

We can use maximum likelihood to provide an estimator of the vector of unknown parameters in the mixture of t -factor analyzers model specified by (5.127) and (5.128). We use the AECM algorithm as outlined in Section 5.14 for mixtures of factor analyzers. The results as outlined in McLachlan and Peel (2000a, Section 3.8) on the consistency of the ML estimator in the case of normal mixture components should carry over here if the adopted factor analysis model holds true for the component distributions.

More specifically, we declare the missing data to be the component-indicators z_{ij} , the factors \mathbf{a}_{ij} in (5.129), and the weights u_j in the characterization (5.134) of the t -distribution for the i th component distribution of \mathbf{W}_j and \mathbf{a}_{ij} . We have from (5.129) and (5.136) that

$$\mathbf{W}_{ij} | \mathbf{a}_{ij}, u_j, z_{ij} = 1 \sim N_p(\boldsymbol{\mu}_i + \mathbf{B}_i \mathbf{a}_{ij}, \mathbf{D}_i/u_j) \quad (5.139)$$

and

$$U_j | z_{ij} = 1 \sim \text{gamma}(\frac{1}{2}\nu_i, \frac{1}{2}\nu_i). \quad (5.140)$$

for $i = 1, \dots, g$.

In the EM framework for this problem, the complete data consist, in addition to the observed data \mathbf{w}_j , of the component-indicators z_{ij} , the unobservable weights w_j , and the latent factors \mathbf{a}_{ij} . The complete-data likelihood $L_c(\Psi)$ can be factored into the product of the marginal densities of the Z_j , the conditional densities of the U_j given the z_j , and the conditional densities of the \mathbf{W}_j given the u_j and the z_j . Accordingly, the complete-data log likelihood can be written as

$$\log L_c(\Psi) = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log Q_{ij} \quad (5.141)$$

where

$$Q_{ij} = \pi_i f_G(u_j; \frac{1}{2}\nu_i, \frac{1}{2}\nu_i) \phi(\mathbf{a}_{ij}; \mathbf{0}, \mathbf{I}_q/u_j) \phi(\mathbf{w}_j; \boldsymbol{\mu}_i + \mathbf{B}_i \mathbf{a}_{ij}, \mathbf{D}_i/u_j), \quad (5.142)$$

and $f_G(u_j; \frac{1}{2}\nu_i, \frac{1}{2}\nu_i)$ denotes the gamma density.

Now $\log Q_{ij}$ can be expressed as

$$\log Q_{ij} = \sum_{h=1}^4 Q_{hij}, \quad (5.143)$$

where

$$Q_{1ij} = \log \pi_i, \quad (5.144)$$

$$\begin{aligned} Q_{2ij} &= -\log \Gamma(\frac{1}{2}\nu_i) + \frac{1}{2}\nu_i \log(\frac{1}{2}\nu_i) \\ &\quad + \frac{1}{2}\nu_i(\log u_j - u_j) - \log u_j, \end{aligned} \quad (5.145)$$

$$Q_{3ij} = -\frac{1}{2}q \log(2\pi) + \frac{1}{2}q \log u_j - \frac{1}{2}\mathbf{a}_{ij}^T \mathbf{a}_{ij}/u_j, \quad (5.146)$$

and

$$\begin{aligned} Q_{4ij} &= -\frac{1}{2}p \log(2\pi) + \frac{1}{2}p \log u_j - \frac{1}{2} \log |\mathbf{D}_i| \\ &\quad - \frac{1}{2}u_j(\mathbf{w}_j - \boldsymbol{\mu}_i - \mathbf{B}_i \mathbf{a}_{ij})^T \mathbf{D}_i^{-1} (\mathbf{w}_j - \boldsymbol{\mu}_i - \mathbf{B}_i \mathbf{a}_{ij}). \end{aligned} \quad (5.147)$$

5.14.5 E-step

It can be seen from (5.147) that in order to carry out the E-step, we need to be able to calculate the conditional expectation of terms like

$$E(Z_{ij} U_j \mathbf{a}_{ij} \mid \mathbf{w}_j), \quad (5.148)$$

$$E(Z_{ij} U_j \mathbf{a}_{ij} \mathbf{a}_{ij}^T \mid \mathbf{w}_j), \quad (5.149)$$

and

$$E(\log U_j \mid \mathbf{w}_j, \mathbf{z}_j) \quad (5.150)$$

for $i = 1, \dots, g$; $j = 1, \dots, n$.

It follows that

$$E_{\Psi^{(k)}}(Z_{ij} \mid \mathbf{w}_j) = \tau_i(\mathbf{w}_j; \Psi^{(k)}), \quad (5.151)$$

where

$$\tau_i(\mathbf{w}_j; \Psi^{(k)}) = \frac{\pi_i^{(k)} f(\mathbf{w}_j; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}, \nu_i^{(k)})}{f(\mathbf{w}_j; \Psi^{(k)})} \quad (5.152)$$

is the posterior probability that \mathbf{w}_j belongs to the i th component of the mixture, using the current fit $\Psi^{(k)}$ for Ψ ($i = 1, \dots, g$; $j = 1, \dots, n$).

We have from (2.44) that

$$U_j \mid \mathbf{w}_j, z_{ij} = 1 \sim \text{gamma}(m_{1i}, m_{2i}), \quad (5.153)$$

where

$$m_{1i} = \frac{1}{2}(\nu_i + p)$$

and

$$m_{2i} = \frac{1}{2}\{\nu_i + \delta(\mathbf{w}_j, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i)\}. \quad (5.154)$$

From (2.46), we have that

$$E(U_j \mid \mathbf{w}_j, z_{ij} = 1) = \frac{\nu_i + p}{\nu_i + \delta(\mathbf{w}_j, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i)}. \quad (5.155)$$

Thus

$$E_{\Psi^{(k)}}(U_j \mid \mathbf{w}_j, z_{ij} = 1) = u_{ij}^{(k)}, \quad (5.156)$$

$$u_{ij}^{(k)} = \frac{\nu_i^{(k)} + p}{\nu_i^{(k)} + \delta(\mathbf{w}_j, \boldsymbol{\mu}_i^{(k)}; \boldsymbol{\Sigma}_i^{(k)})}. \quad (5.157)$$

From (5.64), we have

$$E_{\Psi^{(k)}}(\log U_j | \mathbf{w}_j, z_{ij} = 1) = \log u_{ij}^{(k)} + \{\psi\left(\frac{\nu_i^{(k)} + p}{2}\right) - \log\left(\frac{\nu_i^{(k)} + p}{2}\right)\} \quad (5.158)$$

for $j = 1, \dots, n$.

With the above results, we can complete the E-step, proceeding similarly as in the previous case of normal component factor analyzers.

5.14.6 CM-steps

We use two CM steps in the AECM algorithm, which correspond to the partition of Ψ into the two subvectors Ψ_1 and Ψ_2 , where Ψ_1 contains the mixing proportions, the elements of the $\boldsymbol{\mu}_i$, and the degrees of freedom ν_i ($i = 1, \dots, g$). The subvector Ψ_2 contains the elements of the matrix \mathbf{B}_i of factor loadings and of the diagonal matrix \mathbf{D}_i . On the first cycle, we specify the missing data to be the component-indicator variables Z_{ij} and the weights u_j in the characterization (5.130) of the t -distribution for the component distribution of \mathbf{w}_j . On the $(k+1)$ th iteration of the algorithm, we update the estimates of the mixing proportions using (5.117), where now the posterior probabilities are calculated using the t -density in place of the normal in (5.116). The updated estimate of the i th component mean $\boldsymbol{\mu}_i$ is given by

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^n \tau_i(\mathbf{w}_j; \Psi^{(k)}) u_{ij}^{(k)} \mathbf{w}_j / \sum_{j=1}^n \tau_i(\mathbf{w}_j; \Psi^{(k)}) u_{ij}^{(k)}, \quad (5.159)$$

where the current weight $u_{ij}^{(k)}$ is given by (5.157).

The updated estimate $\nu_i^{(k+1)}$ of ν_i does not exist in closed form, but is given as a solution of the equation

$$\begin{aligned} & \{-\psi\left(\frac{1}{2}\nu_i\right) + \log\left(\frac{1}{2}\nu_i\right) + 1 + \frac{1}{n_i^{(k)}} \sum_{j=1}^n \tau_i(\mathbf{w}_j; \Psi^{(k)}) (\log w_{ij}^{(k)} - w_{ij}^{(k)}) \\ & + \psi\left(\frac{\nu_i^{(k)} + p}{2}\right) - \log\left(\frac{\nu_i^{(k)} + p}{2}\right)\} = 0, \end{aligned} \quad (5.160)$$

where $n_i^{(k)} = \sum_{j=1}^n \tau_i(\mathbf{w}_j; \Psi^{(k)})$ ($i = 1, \dots, g$), and $\psi(\cdot)$ is the Digamma function.

The estimate of Ψ is updated so that its current value after the first cycle is given by

$$\Psi^{(k+1/2)} = (\Psi_1^{(k+1)^T}, \Psi_2^{(k)^T})^T. \quad (5.161)$$

On the second cycle of this iteration, the complete data are expanded to include the unobservable factors \mathbf{U}_{ij} associated with the \mathbf{w}_j . The estimates of the matrix of factor loadings \mathbf{B}_i and the diagonal matrix \mathbf{D}_i can be updated using (5.119) to (5.124), but where the i th component sample covariance matrix is calculated as

$$\mathbf{V}_i^{(k+1/2)} = \frac{\sum_{j=1}^n \tau_i(\mathbf{u}_j; \Psi^{(k+1/2)}) u_{ij}^{(k+1/2)} (\mathbf{u}_j - \boldsymbol{\mu}_i^{(k+1)}) (\mathbf{u}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^n \tau_i(\mathbf{u}_j; \Psi^{(k+1/2)})}, \quad (5.162)$$

where $u_{ij}^{(k+1/2)}$ is updated partially by using $\Psi^{(k+1/2)}$ for Ψ in (5.157).

5.15 PARAMETER-EXPANDED EM (PX-EM) ALGORITHM

In contrast to the AECM algorithm where the optimal value of the working parameter is determined before EM iterations, a variant is considered by Liu et al. (1998) which maximizes the complete-data log likelihood as a function of the working parameter within each EM iteration. The so-called parameter-expanded EM (PX-EM) algorithm has been used for fast stable computation of MLE in a wide range of models.

In this algorithm, the given model is embedded in a larger model with an additional parameter α . Let the original model parameter be Ψ and the parameter vector of the expanded model be $\Xi = (\Psi^*, \alpha)^T$. We let $g_X(\mathbf{y}; \Xi)$ denote the density of the observed-data vector \mathbf{y} under the expanded model. The following conditions are imposed.

1. Ψ^* is of the same dimension as Ψ ;
2. $\Psi = R(\Psi^*, \alpha)$ for a known function R ;
3. when $\alpha = \alpha_0$, $\Psi^* = \Psi$.
4. $g_X(\mathbf{y}; \Psi^*, \alpha) = g_X(\mathbf{y}; \Psi^*, \alpha') \quad \forall \alpha, \alpha'$.

This condition means that the observed data \mathbf{y} carries no information on the parameter α .

5. The parameters in Ξ are identifiable from the complete data \mathbf{x} .

The PX-EM algorithm is the same as the plain EM algorithm to the expanded model. It shares the same convergence properties as the plain EM. Its advantage is a certain amount of gain in speed of convergence if the expansion is appropriately chosen, especially in multivariate incomplete data.

Little and Rubin (2002) formulate a PX-EM algorithm for the estimation of univariate t parameters when the degrees of freedom is known. For definition of the t distribution, see Section 2.6 and Section 5.8.1, where we deal with the multivariate t distribution. The observed data vector $\mathbf{y} = (y_1, \dots, y_n)^T$ is a random sample from a univariate t distribution with known degrees of freedom ν and unknown parameters μ and σ^2 . The expanded data vector consists of n i.i.d. observations from the distribution of (\mathbf{Y}, \mathbf{W}) with the following model with an additional parameter α :

$$(Y | \mu^*, \sigma^*, \alpha; w) \sim N(\mu^*, \sigma^*/w), \quad (W | \mu^*, \sigma^*, \alpha) \sim \alpha \chi_\nu^2 / \nu. \quad (5.163)$$

It can be seen that the expanded model reduces to the original t model when $\alpha = 1$. It can also be seen that the observed data does not have any information on α since the marginal distribution of Y does not involve α ; it can also be noted that the expanded data variable can identify α . The function $\mu = \mu^*, \sigma = \sigma^*/\sqrt{\alpha}$ plays the role of R mentioned in condition (2) above. Little and Rubin (2002) work out the PX-EM steps. They argue that the PX-EM converges faster than the EM and point out that it is because the fraction of missing information in the expanded model is smaller than in the original model. (Refer to our discussion of the relation between the fraction of missing information and rate of convergence in Section 3.9.)

This variant has been further developed, known as the one-step-late PX-EM algorithm, to compute MAP or maximum penalized likelihood (MPL) estimates (van Dyk and Tang, 2003). Analogous convergence results hold for the ECME, AECM, and PX-EM algorithms as for the EM and ECM algorithms. More importantly, these algorithms preserve the monotone convergence of the EM algorithm as stated in (3.1); see Foulley and van Dyk (2000) and van Dyk (2000) for details.

5.16 EMS ALGORITHM

Silverman, Jones, Wilson, and Nychka (1990) modify the EM algorithm by introducing a smoothing step (the S-step) at each EM iteration. They called the resulting procedure the EMS algorithm. When applied to the problem of estimating the vector λ of emission intensities for PET/SPECT data, as considered in Section 2.5, the estimate $\lambda_i^{(k+1)}$ is given by

$$\lambda_s^{(k+1)} = \sum_{i=1}^n w_{si} \lambda_i^{(k)} q_i^{-1} \sum_{j=1}^d \left\{ y_j p_{ij} / \sum_{h=1}^n \lambda_h^{(k)} p_{hj} \right\} \quad (s = 1, \dots, n), \quad (5.164)$$

where $w = ((w_{si}))$ is a smoothing matrix. Usually, w is taken to be row-stochastic. If it is doubly stochastic, then the algorithm preserves photon energy and no normalization step is required; see Green (1990b) and Kay (1994).

In partial justification of this procedure, Nychka (1990) shows that an approximation to the EMS algorithm can be viewed as a penalized likelihood approach. Kay (1994) reports some unpublished work in which he has established the convergence of the EMS algorithm and explained its faster rate of convergence than the ordinary EM algorithm.

5.17 ONE-STEP-LATE ALGORITHM

The OSL (One-Step Late) algorithm was first suggested by Green (1990a) in the context of reconstruction of real tomographic data. It was subsequently proposed by Green (1990b) in a general context for finding the MAP estimate or the MPLE.

With the latter solution, the problem in the context of the analysis of PET/SPECT data as considered in Section 2.5, is to maximize the function

$$\log L(\lambda) - \xi K(\lambda), \quad (5.165)$$

where ξ is some smoothing parameter and $K(\lambda)$ is a roughness functional.

Alternatively,

$$\exp\{-\xi K(\lambda)\}$$

can be regarded as proportional to a prior distribution for λ as proposed, for example, by Geman and McClure (1985). In that case, (5.165) represents the log of the posterior density of λ , ignoring a normalizing constant not involving λ .

In applying the EM algorithm to the problem of maximizing (2.32), the E-step is the same as in Section 2.5, in that it effectively requires the calculation of (2.34). But the M-step is now more complicated since $\lambda^{(k+1)}$ must be obtained by solving a set of highly nonlinear, nonlocal equations,

$$\sum_{j=1}^d z_{ij}^{(k)} / \lambda_i - \sum_{j=1}^d p_{ij} - \xi \partial K(\lambda) / \partial \lambda_i = 0 \quad (i = 1, \dots, n), \quad (5.166)$$

where $z_{ij}^{(k)}$ is the conditional expectation of Z_{ij} given the observed data y , using the current fit $\lambda^{(k)}$ for λ . It is given by (2.34).

Green (1990a) proposes to evaluate the partial derivatives in (5.166) at the current estimate $\boldsymbol{\lambda}^{(k)}$, thus yielding

$$\lambda_i^{(k+1)} = \frac{\sum_{j=1}^d z_{ij}^{(k)}}{\sum_{j=1}^d p_{ij} + \xi \partial K(\boldsymbol{\lambda}^{(k)}) / \partial \lambda_i} \quad (i = 1, \dots, n). \quad (5.167)$$

Green (1990b) proves a local convergence result for this OSL algorithm provided that the value of ξ is not too large. Lange (1990) extends this work and, by introducing a line-search at each iteration, is able to produce a global convergence result.

5.18 VARIANCE ESTIMATION FOR PENALIZED EM AND OSL ALGORITHMS

5.18.1 Penalized EM Algorithm

Segal et al. (1994) synthesize the Supplemented EM algorithm of Meng and Rubin (1991) and the OSL algorithm of Green (1990b) for MPL estimation to suggest a method for computing the asymptotic covariance matrix of the MPLE through the EM computations. Let us consider the maximization of

$$\log L(\boldsymbol{\Psi}) - \xi K(\boldsymbol{\Psi})$$

as in equation (5.165). In the EM algorithm for MPL estimation, the E-step is the same as for ML estimation, but on the M-step

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) - \xi K(\boldsymbol{\Psi}) \quad (5.168)$$

is to be maximized.

We let $\mathbf{J}_P(\boldsymbol{\Psi})$ be the mapping induced by the sequence of iterates $\{\boldsymbol{\Psi}^{(k)}\}$, where $\boldsymbol{\Psi}^{(k+1)}$ is obtained by maximization of (5.168). If this is accomplished by equating the derivative of (5.168) to zero, then Green (1990b) shows that the Jacobian $\mathbf{J}_P(\boldsymbol{\Psi})$ of this mapping at $\boldsymbol{\Psi} = \tilde{\boldsymbol{\Psi}}$ is given by

$$\mathbf{J}_P(\tilde{\boldsymbol{\Psi}}) = \{\mathcal{I}_c(\tilde{\boldsymbol{\Psi}}; \mathbf{y}) + \xi \partial^2 \mathbf{K}(\tilde{\boldsymbol{\Psi}}) / \partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T\}^{-1} \mathcal{I}_m(\tilde{\boldsymbol{\Psi}}; \mathbf{y}), \quad (5.169)$$

where $\tilde{\boldsymbol{\Psi}}$ is the MPLE of $\boldsymbol{\Psi}$.

The asymptotic covariance matrix of the MPLE $\tilde{\boldsymbol{\Psi}}$ can be estimated by \mathbf{V} where, corresponding to the use of the inverse of the observed information matrix in the unpenalized case, it is defined as the inverse of the negative of the Hessian of the right-hand side of (5.168) evaluated at $\boldsymbol{\Psi} = \tilde{\boldsymbol{\Psi}}$. Segal et al. (1994) exploited the representation (4.71) of the observed information matrix in the unpenalized situation to express the asymptotic covariance matrix \mathbf{V} of the MPLE $\hat{\boldsymbol{\Psi}}$ as

$$\mathbf{V} = \{\mathbf{I}_d - \mathbf{J}_P(\tilde{\boldsymbol{\Psi}})\}^{-1} \mathbf{V}_c \quad (5.170)$$

$$= \mathbf{V}_c + \Delta \mathbf{V}, \quad (5.171)$$

where

$$\mathbf{V}_c = \{\mathcal{I}_c(\tilde{\boldsymbol{\Psi}}; \mathbf{y}) + \xi \partial^2 \mathbf{K}(\tilde{\boldsymbol{\Psi}}) / \partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T\}^{-1}$$

and

$$\Delta \mathbf{V} = \{\mathbf{I}_d - \mathbf{J}_P(\tilde{\boldsymbol{\Psi}})\}^{-1} \mathbf{J}_P(\tilde{\boldsymbol{\Psi}}) \mathbf{V}_c.$$

The diagonal elements of $\Delta \mathbf{V}$ give the increases in the asymptotic variances of the components of the MPLE $\tilde{\Psi}$ due to not having observed \mathbf{x} fully.

As with the Supplemented EM algorithm, $\mathbf{J}_P(\tilde{\Psi})$ can be computed by numerical differentiation, using the penalized EM code; $\mathcal{I}_c(\tilde{\Psi}; \mathbf{y})$ can be computed by complete-data methods. The term $\partial^2 \mathbf{K}(\tilde{\Psi}) / \partial \Psi \partial \Psi^T$ has to be evaluated analytically except in simple cases like quadratic $K(\Psi)$, when it becomes a constant.

5.18.2 OSL Algorithm

As noted in Section 1.6.3, the MPLE of Ψ can be computed using the OSL algorithm. Green (1990b) has shown that the Jacobian $\mathbf{J}_{OSL}(\Psi)$ of the mapping induced by the OSL algorithm is given at $\Psi = \tilde{\Psi}$ by

$$\mathbf{J}_{OSL}(\tilde{\Psi}) = \mathcal{I}_c^{-1}(\tilde{\Psi}; \mathbf{y}) \{ \mathcal{I}_m(\tilde{\Psi}; \mathbf{y}) - \xi \partial^2 K(\tilde{\Psi}) / \partial \Psi \partial \Psi^T \}. \quad (5.172)$$

Adapting the result (4.73) to the OSL algorithm, Segal et al. (1994) note that the asymptotic covariance matrix of the MPLE $\tilde{\Psi}$ can be expressed as

$$\{ \mathbf{I}_d - \mathbf{J}_{OSL}(\tilde{\Psi}) \}^{-1} \mathcal{I}_c^{-1}(\tilde{\Psi}; \mathbf{y}).$$

It is interesting to note that this formula does not directly involve the penalty term. Similarly, as for the Jacobians of the EM and penalized EM maps, $\mathbf{J}_{OSL}(\tilde{\Psi})$ can be computed by numerical differentiation, using the code for the OSL algorithm.

5.18.3 Example 5.9: Variance of MPLE for the Multinomial (*Examples 1.1 and 4.1 Continued*)

To give a numerical example, we return to Example 4.1, involving the multinomial data (1.12). This illustration is taken from Segal et al. (1994), who used it to demonstrate the use of (5.170) for providing an estimate of the variance of the MPLE obtained via the EM algorithm.

Following Green (1990b), Segal et al. (1994) added the penalty function

$$-10(\Psi - 0.5)^2 \quad (5.173)$$

to the log likelihood function $\log L(\Psi)$ given by (1.14). The use of (5.173) corresponds to a truncated normal prior for Ψ . In this case, the MPLE of Ψ , $\tilde{\Psi}$, is a solution of the equation

$$\frac{y_1}{2 + \Psi} - \frac{y_2 + y_3}{1 - \Psi} + \frac{y_4}{\Psi} - 20(\Psi - 0.5) = 0. \quad (5.174)$$

This equation can be solved directly, being a quartic polynomial in Ψ with relevant root $\tilde{\Psi} = 0.60238108$. On differentiation of the left-hand side of (5.174) and taking the negative inverse evaluated at $\Psi = \tilde{\Psi}$, the estimate of the variance of $\tilde{\Psi}$ so obtained is 0.00256.

We can apply the EM algorithm to this problem in the same manner as in the unpenalized situation in Section 1.4.2. The E-step is the same, effectively requiring the calculation of (1.23) on the $(k+1)$ th iteration. The M-step now takes $\Psi^{(k+1)}$ to be the root of the cubic equation

$$\frac{y_{12}^{(k)}}{2 + \Psi} - \frac{y_2 + y_3}{1 - \Psi} + \frac{y_4}{\Psi} - 20(\Psi - 0.5) = 0. \quad (5.175)$$

Table 5.2 Results of the EM Algorithm for DLR Multinomial Data.

Iteration (k)	$\Psi^{(k)}$	$d^{(k)} = \Psi^{(k)} - \tilde{\Psi}$	$r^{(k)} = d^{(k+1)}/d^{(k)}$
0	0.50000000	-0.12038109	0.143085
1	0.60315651	-0.01722467	0.132367
2	0.61810110	-0.00227998	0.130946
3	0.62008253	-0.00029856	0.130760
4	0.62034205	-0.00003904	0.130736
5	0.62037598	-0.00000510	0.130732
6	0.62038042	-0.00000067	0.130762
7	0.62038100	-0.00000009	0.130732
8	0.62038108	-0.00000001	0.130732

Source: Adapted from Segal et al. (1994).

Details on the convergence of the sequence $\{\Psi^{(k)}\}$ to $\tilde{\Psi}$ are displayed in Table 5.2, along with the ratios $r^{(k)}$ of successive deviations. It can be seen that they are essentially constant for $k \geq 3$ and are accurate to five decimal places at $k = 5$, where $J_P(\tilde{\Psi}) = 0.130732$.

We have seen in Section 4.2.5 that $I_c(\hat{\Psi}; \mathbf{y})$ is given by the binomial variance, so that

$$\begin{aligned} I_c(\tilde{\Psi}; \mathbf{y}) &= \frac{y_{12}^{(\infty)} + y_2 + y_3 + y_4}{\tilde{\Psi}(1 - \tilde{\Psi})} \\ &= 431.38, \end{aligned} \quad (5.176)$$

where

$$y_{12}^{(\infty)} = \frac{1}{4}y_1\tilde{\Psi}/\left(\frac{1}{2} + \frac{1}{4}\tilde{\Psi}\right).$$

Also, $\xi \partial^2 K(\tilde{\Psi})/\partial \Psi \partial \Psi^T = 20$, and so from (5.170), we finally obtain $V = 0.00255$, which is decomposed according to equation (5.171) as $V_c = 0.00222$ and $\Delta V = 0.00033$.

5.19 INCREMENTAL EM

Neal and Hinton (1998) proposed the incremental EM (IEM) algorithm to improve the convergence rate of the EM algorithm. With this algorithm, the available n observations are divided into $B/(B \leq n)$ blocks and the E-step is implemented for only a block of data at a time before performing a M-step. A “scan” of the IEM algorithm thus consists of B partial E-steps and B M-steps. The argument for improved rate of convergence is that the algorithm exploits new information more quickly rather than waiting for a complete scan of the data before parameters are updated by an M-step. Another method suggested by Neal and Hinton (1998) is the sparse EM (SPEM) algorithm. In fitting a mixture model to a data set by ML via the EM, the current estimates of some posterior probabilities $\tau_{ij}^{(k)}$ for a given data point \mathbf{y}_j are often close to zero. For example, if $\tau_{ij}^{(k)} < 0.005$ for the first two components of a four-component mixture being fitted, then with the SPEM algorithm we would fix $\tau_{ij}^{(k)}$ ($i = 1, 2$) for membership of \mathbf{y}_j with respect to the first two components at their current values and only update $\tau_{ij}^{(k)}$ ($i = 3, 4$) for the last two components. This sparse E-step will take time proportional to the number of components that needed to be updated.

A sparse version of the IEM algorithm (SPIEM) can be formulated by combining the partial E-step and the sparse E-step. With these versions, the likelihood is still increased after each scan. Ng and McLachlan (2003a) study the relative performances of these algorithms with various number of blocks B for the fitting of normal mixtures. They propose to choose B to be that factor of n that is the closest to $B^* = \text{round}(n^{2/5})$ for unrestricted component-covariance matrices, where $\text{round}(r)$ rounds r to the nearest integer.

Other approaches for speeding up the EM algorithm for mixtures have been considered in Bradley, Fayyad, and Reina (1998) and Moore (1999). The former developed a scalable version of the EM algorithm to handle very large databases with a limited memory buffer. It is based on identifying regions of the data that are compressible and regions that must be maintained in memory. Moore (1999) has made use of multiresolution kd-trees (*mrkd-trees*) to speed up the fitting process of the EM algorithm on normal mixtures. Here *kd* stands for k -dimensional where, in our notation, $k = p$, the dimension of an observation y_j . His approach builds a multiresolution data structure to summarize the database at all resolutions of interest simultaneously. The *mrkd-tree* is a binary tree that recursively splits the whole set of data points into partitions. The contribution of all the data points in a tree node to the sufficient statistics is simplified by calculating at the mean of these data points to save time. Ng and McLachlan (2003c) combined the IEM algorithm with the *mrkd-tree* approach to further speed up the EM algorithm. They also studied the convergence properties of this modified version and the relative performance with some other variants of the EM algorithm for speeding up the convergence for the fitting of normal mixtures. Ng and McLachlan (2004b) proposed to speed up the SPIEM algorithm further by imposing a multiresolution *kd-tree* structure in performing the E-step. They also considered a second version that involves “pruning” the tree nodes. They showed that these two new SPIEM multiresolution *kd-tree*-based algorithms provide a fast EM-based mixture model approach to the segmentation of three-dimensional magnetic resonance images.

Neither the scalable EM algorithm nor the *mrkd-tree* approach guarantees the desirable reliable convergence properties of the EM algorithm. Moreover, the scalable EM algorithm becomes less efficient when the number of components g is large, and the *mrkd-trees*-based algorithms slow down as the dimension p increases; see, for example, Ng and McLachlan (2003a) and the references therein.

5.20 LINEAR INVERSE PROBLEMS

Vardi and Lee (1993) note that many problems in science and technology can be posed as linear inverse problems with positivity restrictions, which they referred to as LININPOS problems. They show that all such problems can be viewed as statistical estimation problems from incomplete data based on “infinitely large samples”, which could be solved by ML method *via* the EM algorithm. In the framework of Vardi and Lee (1993), LININPOS problems require solving the equation

$$g(y) = \int_{D_{g_c}} h(x, y) g_c(x) dx, \quad (5.177)$$

where D_{g_c} and D_g are the domains of the nonnegative real-valued functions g_c and g , respectively. In image analysis, g_c represents the true distorted image that would have been recorded, had there been no blurring in the image recording process, and g represents the recorded blurred image. The values of g_c and g are grey-level intensities. The function

$h(x, y)$, which is assumed to be a bounded nonnegative function on $D_{g_c} \times D_g$, characterizes the blurring mechanism.

As noted by Vardi and Lee (1993), the class of LININPOS problems and the corresponding class of estimation problems based on incomplete data are rich and include examples for which the EM algorithm has been independently derived. Examples include image reconstruction in PET/SPECT, as discussed in Section 2.5, as well as more traditional statistical estimation problems, such as ML estimation from grouped and truncated data, as considered in Section 2.8.

MONTE CARLO VERSIONS OF THE EM ALGORITHM

6.1 INTRODUCTION

In the last two decades or so a large body of methods has emerged based on iterative simulation techniques useful especially in computing Bayesian solutions to the kind of incomplete-data problems discussed in the previous chapters. Most of these methods are aimed at estimating the entire posterior density and not just finding the maximum *a posteriori* (MAP) estimate of the parameter vector Ψ . As emphasized in Gelman and Rubin (1992), it is wise to find MLE or MAP estimates of Ψ before using iterative simulation, because of the difficulties in assessing convergence of iterative simulation methods, particularly when the regions of high density are not known *a priori*. In many problems where such iterative simulation or Monte Carlo techniques are used, the EM algorithm often comes in handy to provide these MLE or MAP estimates. Although this establishes a connection between EM and Monte Carlo methods, there are stronger and more significant connections, which we explore in this chapter.

These iterative simulation techniques are conceptually very similar, simply replacing the E- and M-steps by draws from the current conditional distribution of the missing data and Ψ , respectively. However, in some methods such as the Monte Carlo EM (MCEM) algorithm, which is one of the simpler extensions of the EM algorithm, only the E-step is so implemented. Many of these methods can be interpreted as iterative simulation analogs of the various versions of the EM algorithm and its extensions.

A Monte Carlo approach is often followed in either or both of the two steps (E and M) of the EM algorithm if the step is analytically intractable. In some situations it is implemented

via a Markov chain Monte Carlo (MCMC) method. In some Bayesian problems, the computation of the posterior mode or the marginal posterior mode (a maximization problem) is facilitated by an EM algorithm. The two-stage structure of the EM algorithm involving a latent variable (missing data) and data augmentation is analogous to the Gibbs sampling algorithm used for the Bayesian version of the same problem. After a brief introduction to various Monte Carlo and Markov chain Monte Carlo (MCMC) algorithms, we give a variety of examples to illustrate these Monte Carlo approaches. Some of these algorithms, such as Rejection Sampling, are not based on Markov chains; neither are they direct methods for random sample generation. We call them ‘Independent and Identically Distributed (i.i.d.) Monte Carlo’ algorithms. We also describe the strong connection between EM algorithm and Gibbs sampling; this includes a connection in their convergence properties.

6.2 MONTE CARLO TECHNIQUES

6.2.1 Integration and Optimization

Integration and optimization are two important mathematical devices used in the development and execution of statistical techniques. Finding unbiased estimators, posterior expectations and moments, and similar exercises need integration. Least squares, maximum likelihood, minimum variance, maximum a posteriori and similar criteria need optimization methods. Thus both the frequentist and Bayesian paradigms need to take recourse to these two techniques. Mostly, though not exclusively, one encounters optimization endeavors in maximum likelihood tasks and integration endeavors in solving Bayesian problems. Depending upon the complexity of the models and the criteria formulated, these integration and optimization exercises may be analytically intractable. Two common approaches to solve such analytically intractable problems are numerical methods and Monte Carlo methods.

As we have noted repeatedly in the book so far, these two devices or operations of integration and optimization form the crucial aspects of the EM algorithm—the E-step is an expectation (integration) computation, and the M-step is an optimization computation. Thus if one or both of these steps are analytically intractable in an EM algorithm, then numerical methods or Monte Carlo methods may need to be invoked. Thus a complicated and analytically intractable E-step may be facilitated by Newton-Cotes and similar numerical quadrature methods, or by Monte Carlo integration, which replaces an integral by a finite sum based on appropriate random samples.

Typical analytically intractable expectation problems in statistical tasks are high-dimensional, as frequently happens in multiparameter Bayesian problems. It is by now well established that if the dimension of the integral is large then quadrature methods perform poorly in comparison to Monte Carlo methods, since the former methods need a large number of points. Moreover, quadrature methods do not perform well unless the functions are smooth. On the other hand, Monte Carlo sample size does not depend as much on dimensionality as quadrature methods do; however, error declines more slowly for Monte Carlo than for quadrature with sample size. But besides affordability of samples, other variance reduction techniques like importance sampling can be brought to bear in Monte Carlo work to offset the slow error rate decline. Lange (1999) presents a lucid discussion of these issues.

In Monte Carlo methods, a random sample w_1, \dots, w_n is drawn from the distribution of a (continuous) random variable W appropriate to the problem. For an arbitrary function

a , the classical Monte Carlo integration estimator of the expectation of $a(W)$ is

$$\hat{E}\{a(W)\} = \frac{1}{n} \sum_{j=1}^n a(w_j).$$

The strong law of large numbers guarantees convergence of $\hat{E}\{a(W)\}$ almost surely to the required expectation $E\{a(W)\}$. An estimate of the variance of the estimate is given by

$$\text{var}[\hat{E}\{a(W)\}] = \frac{1}{n(n-1)} \sum_{j=1}^n [a(w_j) - \hat{E}\{a(W)\}]^2.$$

A Monte Carlo estimate is subject not only to random variation within a Monte Carlo sample, but also to random variation from one possible Monte Carlo sample to another. The error due to the latter part is known as Monte Carlo error. In a Monte Carlo exercise it is important to estimate the Monte Carlo error as well.

Monte Carlo optimization techniques perform random explorations of the surface over which the optimum is to be found. If done carefully, one can avoid being trapped in local extrema and move towards global extrema. In addition to helping to locate the optimum, Monte Carlo methods allow one to explore a function and find an approximation to it.

6.2.2 Example 6.1: Monte Carlo Integration

Consider the evaluation of the integral

$$J = \int_0^1 \cos\left(\frac{\pi w}{2}\right) dw.$$

Actually, this is easily evaluated to be $\frac{2}{\pi} \approx \frac{7}{11} \approx 0.63636$. Suppose we wish to evaluate the integral by Monte Carlo methods. Following the above discussion, we can use the estimate

$$J_n = \frac{1}{n} \sum_{j=1}^n \cos\left(\frac{\pi w_j}{2}\right)$$

where the w_j are drawn from the uniform distribution on $[0, 1]$. We did this for $n = 10,000$ with the results and statistics shown in Table 6.1 and in Figure 6.1.

6.3 MONTE CARLO EM

6.3.1 Introduction

In an EM algorithm, the E-step may be difficult to implement because of difficulty in computing the expectation of log likelihood. Wei and Tanner (1990a, 1990b) suggest a Monte Carlo approach by simulating the missing data Z from the conditional distribution $k(z | y; \Psi^{(k)})$ on the E-step of the $(k + 1)$ th iteration, then maximizing the approximate conditional expectation of the complete-data log likelihood

$$\hat{Q}(\Psi; \Psi^{(k)}) = \frac{1}{m} \sum_{j=1}^m \log L_c(\Psi; y, z_j). \quad (6.1)$$

Table 6.1 Descriptive Statistics of Uniform [0, 1] (W) and $V = \cos(\pi W/2)$ from 10,000 Samples of W .

Statistic	W	Theoretical Values	V	Theoretical Values
No. of cases	10,000	—	10,000	—
Minimum	0.00006	0	0.00015	0
Maximum	0.99	1	1.00000	1
Median	0.50459	0.5	0.70176	0.70711
Mean	0.50210	0.5	0.63374	0.63636
Standard Dev	0.28962	0.28868	0.30940	0.30822
Variance	0.08388	0.08333	0.09573	0.095

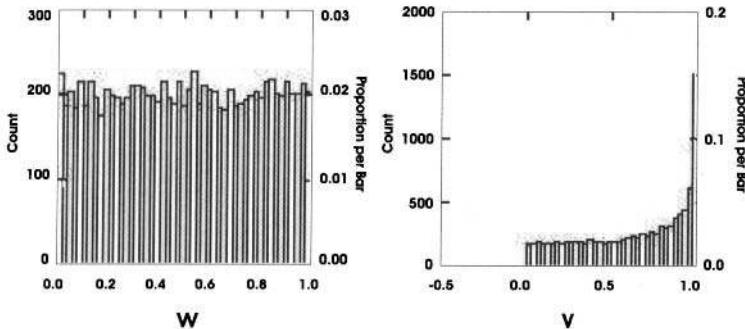


Figure 6.1 (a) Histogram of Uniform [0,1] (W); (b) Histogram of $V = \cos(\pi W/2)$ from 10,000 Samples of W .

The limiting form of this as $m \rightarrow \infty$ is the actual $Q(\Psi; \Psi^{(k)})$. This is exactly the idea of Monte Carlo integration discussed earlier. Although the maximization of (6.1) can often be difficult, sometimes, as in exponential family situations, there can be closed-form solutions to the maximization problem.

To reduce costs of Monte Carlo sampling in MCEM, one may reuse samples from previous expectation steps (Levine and Casella, 2001). When it is difficult to simulate the E-step, Gibbs sampling is often useful (Chan and Ledolter, 1995) as will be seen later in this chapter. In view of the randomness and variability involved in Monte Carlo results, it is generally recommended that convergence criteria for Monte Carlo optimization methods be tighter than for deterministic optimization methods.

In MCEM, a Monte Carlo error is introduced at the E-step and the monotonicity property is lost. But in certain cases, the algorithm gets close to a maximizer with a high probability (Booth and Hobert, 1999). The problems of specifying m and monitoring convergence are of central importance in the routine use of the algorithm. Wei and Tanner (1990a) recommend that small values of m be used in the initial stages and values be increased as the algorithm moves closer to convergence. As to monitoring convergence, they recommend that the

values of $\Psi^{(k)}$ be plotted against k and if convergence is indicated by the stabilization of the process with random fluctuations about $\hat{\Psi}$, the process may be terminated; otherwise, the process be continued with a larger value of m . Alternative schemes for specifying m and stopping rules are considered by Booth and Hobert (1999) and McCulloch (1997).

We give below two examples of MCEM. For an example of the application of MCMC and Monte Carlo EM in multiple change point estimation, see Chib (1998).

6.3.2 Example 6.2: Monte Carlo EM for Censored Data from Normal

This example is from Robert and Casella (2004). Suppose w_1, \dots, w_n is a random sample from $N(\mu, 1)$. Let the observations be in increasing order such that w_1, \dots, w_m are uncensored and w_{m+1}, \dots, w_n are censored at c (that is, we only know that they are $\geq c$, but not their actual values). Let $\mathbf{z} = (w_1, \dots, w_m, w_{m+1}, \dots, w_n)^T$ denote the complete-data vector and

$$\mathbf{z} = (w_{m+1}, \dots, w_n)^T$$

be the vector containing the missing data. Also, let \bar{w} be the mean of the m uncensored observations. The complete-data log likelihood function for $\Psi = \mu$ (apart for an additive constant not involving μ) is

$$\log L_c(\mu) = -\sum_{j=1}^m \frac{(w_j - \mu)^2}{2} - \sum_{j=m+1}^n \frac{(w_j - \mu)^2}{2}.$$

The density of the missing data \mathbf{z} is a product of truncated normals and so

$$\log k(\mathbf{z} | \mathbf{y}; \mu) \propto \sum_{j=m+1}^n \frac{(w_j - \mu)^2}{2}.$$

On the E-step, the expected conditional complete-data log likelihood is calculated to give (but for an additive constant not involving μ)

$$Q(\mu; \mu^{(k)}) = -\frac{1}{2} \left[\sum_{j=1}^m (w_j - \mu)^2 + \sum_{j=m+1}^n E_{\mu^{(k)}} \{ (W_j - \mu)^2 | W_j > c \} \right],$$

leading on the M-step to the updated estimate of μ ,

$$\mu^{(k+1)} = \frac{m\bar{w} + (n-m)E_{\mu^{(k)}}(W | W > c)}{n}. \quad (6.2)$$

On substituting for the current conditional expectation of W in (6.2), we obtain

$$\mu^{(k+1)} = \frac{m}{n}\bar{w} + \frac{n-m}{n}\mu^{(k)} + \frac{1}{n} \frac{\phi(c - \mu^{(k)})}{1 - \Phi(c - \mu^{(k)})},$$

where ϕ is the standard normal density function and Φ is the standard normal distribution function.

The MCEM solution in this example is to replace $E_{\mu^{(k)}}(W | W > c)$ by

$$\frac{1}{m} \sum_{j=1}^m w_j,$$

where w_j is generated from the truncated (at c) normal density with mean $\mu^{(k)}$ and unit variance.

6.3.3 Example 6.3: MCEM for a Two-Parameter Multinomial (*Example 2.4 Continued*)

In Example 2.4, if we were to employ an MC E-step, we could draw z_{11}, \dots, z_{1m} and z_{21}, \dots, z_{2m} , respectively, from independent binomial distributions with sample size n_A and probability parameter

$$p^{(k)^2} / (p^{(k)^2} + 2p^{(k)}r^{(k)}),$$

and with sample size n_B and probability parameter

$$q^{(k)^2} / (q^{(k)^2} + 2q^{(k)}r^{(k)}),$$

with $\Psi^{(k)}$ used in place of the unknown parameter vector Ψ on the $(k+1)$ th iteration. These could then be used instead of equation (2.27) as

$$n_{AA}^{(k)} = \bar{z}_{1m} = \frac{1}{m} \sum_{j=1}^m z_{1i}, \quad n_{BB}^{(k)} = \bar{z}_{2m} = \frac{1}{m} \sum_{j=1}^m z_{2i}. \quad (6.3)$$

6.3.4 MCEM in Generalized Linear Mixed Models

In the generalized mixed effects model, the EM algorithm gives rise to difficulties because the E-step is intractable even under Gaussian assumptions of random effects. Steele (1996) develops an approximation to the E-step using Laplace's method; although conceptually simple, the method does not guarantee the usual convergence properties of the EM algorithm. Booth and Hobert (1999) approximate the E-step using an independent sampler based on multivariate importance sampling or rejection sampling; see Sections 6.6.3 and 6.6.2.1. They also implement a convergence monitoring method within the E-step. Chen, Zhang, and Davidian (2002) relax the Gaussian assumption and require only that the random effects belong to a class of 'smooth' densities, which may be skewed, multi-modal, fat- or thin-tailed *vis à vis* the normal; they approximate the density by a semiparametric approach, use an efficient algorithm to sample from the semiparametric density and use a Monte Carlo EM algorithm to estimate fixed parameters, variance components and the density. Vaida et al. (2004) deal with generalized linear mixed effects models (GLMM) and Proportional Hazards Mixed Effects Models (PHMM), where the marginal likelihoods are intractable and they show how to compute the E-step by a Monte Carlo simulation. Vaida and Meng (2005) consider GLMM with binary response and develop a two-slice EM algorithm where by further augmenting the random effects (considered to be missing data) used in the M-step, they implement the Monte Carlo E-step via a slice sampler, a Markov chain Monte Carlo technique.

We let $\mathbf{w} = (w_1, \dots, w_n)^T$ denote the observed data vector. Conditional on the unobservable random effects vector, $\mathbf{u} = (u_1, \dots, u_q)^T$, we assume that \mathbf{w} arises from a GLM. The conditional mean $\mu_j = E(y_j | \mathbf{u})$ is related to the linear predictor $\eta_j = \mathbf{x}_j^T \boldsymbol{\beta} + \mathbf{z}_j^T \mathbf{u}$ by the link function $h(\mu_j) = \eta_j$ ($j = 1, \dots, n$), where $\boldsymbol{\beta}$ is a p -vector of fixed effects and \mathbf{x}_j and \mathbf{z}_j are, respectively, a p -vector and q -vector of explanatory variables associated with the fixed and random effects. This formulation encompasses the modeling of data involving multiple sources of random error, such as repeated measures within subjects and clustered data collected from some experimental units (Breslow and Clayton, 1993).

We let the distribution for \mathbf{u} be $p(\mathbf{u}; \mathbf{D})$ that depends on the parameter vector \mathbf{D} . The observed data y_j are conditionally independent with density functions of the form

$$f(y_j | \mathbf{u}; \boldsymbol{\beta}, \kappa) = \exp[m_j \kappa^{-1} \{\theta_j y_j - b(\theta_j)\} + c(y_j; \kappa)], \quad (6.4)$$

where θ_j is the canonical parameter, κ is the dispersion parameter, m_j is the known prior weight, and b and c are scalar-valued functions. The conditional mean and canonical parameters are related through the equation $\mu_j = \partial b(\theta_j)/\partial\theta_j$. Let Ψ denote the vector of unknown parameters within β , κ , and D . The likelihood function for Ψ is given by

$$L(\Psi) = \int \prod_{j=1}^n f(y_j | \mathbf{u}; \beta, \kappa) p(\mathbf{u}; D) d\mathbf{u}, \quad (6.5)$$

which cannot usually be evaluated in closed form due to an intractable integral whose dimension depends on the structure of the random effects.

Within the EM framework, the random effects are considered as missing data. The complete-data vector is then $x = (\mathbf{y}^T, \mathbf{u}^T)^T$ and the complete-data log likelihood function is given by

$$\log L_c(\Psi) = \sum_{j=1}^n \log f(y_j | \mathbf{u}; \beta, \kappa) + \log p(\mathbf{u}; D). \quad (6.6)$$

On the $(k+1)$ th iteration of the EM algorithm, the E-step involves the computation of the Q-function, $Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{\log L_c(\Psi) | \mathbf{y}\}$, where the expectation is with respect to the conditional distribution of $\mathbf{u} | \mathbf{y}$ with current parameter value $\Psi^{(k)}$. As this conditional distribution involves the (marginal) likelihood function $L(\Psi)$ given in (6.5), an analytical evaluation of the Q-function for the model (6.4) will be impossible outside the normal theory mixed model (Booth and Hobert, 1999). The MCEM algorithm can be adopted to tackle this problem by replacing the expectation in the E-step with a Monte Carlo approximation. Let $\mathbf{u}^{(1_k)}, \dots, \mathbf{u}^{(M_k)}$ denote a random sample from $k(\mathbf{u} | \mathbf{y}; \Psi^{(k)})$ at the $(k+1)$ th iteration. An MC approximation of the Q-function is given by

$$Q_M(\Psi; \Psi^{(k)}) = \frac{1}{M} \sum_{m=1}^M \left\{ \log f(\mathbf{y} | \mathbf{u}^{(m_k)}; \beta, \kappa) + \log p(\mathbf{u}^{(m_k)}; D) \right\}. \quad (6.7)$$

From (6.7), it can be seen that the first term of the approximated Q-function involves only parameters β and κ , while the second term involves only D . Thus, the maximization in the MC M-step is usually relatively simple within the GLMM context (McCulloch, 1997).

Alternative simulation schemes for \mathbf{u} can be used for (6.7). For example, Booth and Hobert (1999) proposed rejection sampling and multivariate t importance sampling approximations. McCulloch (1997) considered dependent MC samples using the MC Newton-Raphson (MCNR) algorithm.

6.3.5 Estimation of Standard Error with MCEM

As observed earlier, the EM algorithm does not produce standard errors of estimates in a natural way. Many methods have been suggested for the computation of standard errors in the EM context. Among these, the one best suited to be adapted for MCEM is Louis' method, as considered in Section 4.7.2; see also Robert and Casella (2004). We use here our standard notations for the score statistic, and observed and expected information matrices for the incomplete and complete data.

An estimate of the covariance matrix of the MLE $\hat{\Psi}$ is given by the inverse of the observed information matrix $I(\hat{\Psi}; \mathbf{y})$. As discussed in Section 4.7.2, Louis (1982) has provided a formula for the computation of $I(\Psi; \mathbf{y})$ in terms of the expectation of first and second derivatives of the complete-data log likelihood function. From (4.93),

$$I(\hat{\Psi}; \mathbf{y}) = \mathcal{I}_c(\hat{\Psi}; \mathbf{y}) - [\text{cov}_{\Psi}\{S_c(\mathbf{X}; \Psi) | \mathbf{y}\}]_{\Psi=\hat{\Psi}}, \quad (6.8)$$

where

$$\mathcal{I}_c(\Psi) = E_{\Psi}\{-\partial^2 \log L_c(\Psi)/\partial \Psi \partial \Psi^T\}.$$

We consider now the case of a single unknown parameter Ψ and write the complete-data log likelihood function $\log L_c(\Psi)$ as $\log L_c(\Psi; \mathbf{y}, \mathbf{z})$. With the view to calculating the expectations in (6.8) by Monte Carlo evaluation, we can express $I(\Psi; \mathbf{y})$ in the form

$$\begin{aligned} I(\Psi; \mathbf{y}) &\approx \frac{1}{m} \sum_{j=1}^m -\partial^2 \log L_c(\Psi; \mathbf{y}, \mathbf{z}^{(j)})/\partial \Psi^2 \\ &+ \frac{1}{m} \sum_{j=1}^m \{\partial \log L_c(\Psi; \mathbf{y}, \mathbf{z}^{(j)})/\partial \Psi - \frac{1}{m} \sum_{j=1}^m \partial \log L_c(\Psi; \mathbf{y}, \mathbf{z}^{(j)})/\partial \Psi\}^2, \end{aligned} \quad (6.9)$$

where $\mathbf{z}^{(j)}$ ($j = 1, \dots, m$) are generated from the missing data distribution, using the MCEM estimate of Ψ .

6.3.6 Example 6.4: MCEM Estimate of Standard Error for One-Parameter Multinomial (Example 1.1 Continued)

From (1.22), the complete-data log likelihood (but for an additive constant not involving Ψ) is

$$\log L_c(\Psi; \mathbf{y}, \mathbf{z}) = (y_{12} + y_4) \log \Psi + (y_2 + y_3) \log(1 - \Psi), \quad (6.10)$$

where $\mathbf{z} = y_{12}$ is the missing data.

In this example, as seen in Section 4.2.5, the direct calculation of the observed information is straightforward. However, a Monte Carlo approach would run as follows: On differentiation of (6.10), we have that

$$\partial \log L_c(\Psi)/\partial \Psi = \frac{y_{12} + y_4}{\Psi} - \frac{y_2 + y_3}{1 - \Psi} \quad (6.11)$$

and

$$\partial^2 \log L_c(\Psi)/\partial \Psi^2 = -\frac{y_{12} + y_4}{\Psi^2} - \frac{y_2 + y_3}{(1 - \Psi)^2}. \quad (6.12)$$

From (6.11) and (6.12), we can form (6.9) to calculate an estimate of the standard error of the MCEM estimate, using values of y_{12} generated from its conditional distribution given the observed data, using the MCEM estimate for Ψ . We have seen in Section 4.2.5 that conditional on the observed data, Y_{12} has a binomial distribution,

$$Y_{12} \sim \text{Binomial}(y_1, p), \quad (6.13)$$

where

$$p = \frac{1}{4}\Psi / (\frac{1}{2} + \frac{1}{4}\Psi).$$

In this example, MCEM yielded an estimate of $\hat{\Psi} = 0.6018$, and its standard error was estimated using (6.9) to be 0.0511, which is fairly close to that based on the observed information calculated directly; see Section 4.2.5.

6.3.7 Stochastic EM Algorithm

Prior to the appearance of the MCEM algorithm, Broniatowski, Celeux, and Diebolt (1983), and Celeux and Diebolt (1985, 1986a, 1986b) considered a modified version of the EM algorithm in the context of computing the MLE for finite mixture models. They called it the Stochastic EM algorithm. It is the same as the MCEM algorithm with $m = 1$. We discuss Celeux and Diebolt's method for the mixture problem here.

We let $\boldsymbol{\theta}$ denote the vector containing the parameters known *a priori* to be distinct in the g component densities $f_i(\mathbf{w}; \boldsymbol{\theta}_i)$ of the mixture to be fitted. Then the totality of parameters for the problem is

$$\boldsymbol{\Psi} = (\boldsymbol{\theta}^T, \boldsymbol{\pi}^T)^T,$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{g-1})^T$ and $\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i$. Suppose we have an observed random sample $\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T$ from the mixture. In the application of the EM algorithm to this problem, the missing-data vector \mathbf{z} is taken to be

$$\mathbf{z} = (\mathbf{z}_1^T, \dots, \mathbf{z}_n^T)^T,$$

where \mathbf{z}_j is the vector of zero-one indicator variables that define the component from which the j th observation \mathbf{w}_j arises ($j = 1, \dots, n$). On the E-step of the EM algorithm, each \mathbf{z}_j vector is replaced (because of the linearity of the complete-data log likelihood in \mathbf{z}) by its current conditional expectation given the observed data \mathbf{y} , which is

$$\mathbf{z}_j^{(k)} = (z_{1j}^{(k)}, \dots, z_{gj}^{(k)})^T,$$

where

$$\begin{aligned} z_{ij}^{(k)} &= \tau_i(\mathbf{w}_j; \boldsymbol{\Psi}^{(k)}) \\ &= \pi_i^{(k)} f_i(\mathbf{w}_j; \boldsymbol{\theta}_i^{(k)}) / f(\mathbf{w}_j; \boldsymbol{\Psi}^{(k)}) \end{aligned} \quad (6.14)$$

is the (current) posterior probability that the j th observation arises from the i th component of the mixture, ($i = 1, \dots, g$; $j = 1, \dots, n$).

However, with the Stochastic EM algorithm, the current posterior probabilities are used in a Stochastic E-step, wherein a single draw is made from the current conditional distribution of \mathbf{z} given the observed data \mathbf{y} . Because of the assumption of independence of the complete-data observations, this is effected by conducting a draw for each j ($j = 1, \dots, n$). That is, a draw $\mathbf{z}_j^{(1_k)}$ is made from the multinomial distribution with g categories having probabilities specified by (6.14). This effectively assigns each observation \mathbf{w}_j outright to one of the g components of the mixture. The M-step then consists of finding the MLE of $\boldsymbol{\Psi}$ as if $\mathbf{w}_1, \dots, \mathbf{w}_n$ were deterministically classified according to $\mathbf{z}_1^{(1_k)}, \dots, \mathbf{z}_n^{(1_k)}$. This contrasts with the EM algorithm where these computations are weighted with respect to the components of the mixture according to the current posterior probabilities. Note that with this notation, on the k th iteration, $\mathbf{z}_j^{(k)}$ denotes the current conditional expectation of \mathbf{Z}_j , the random variable associated with \mathbf{z}_j , while $\mathbf{z}_j^{(m_k)}$ denotes the m th random draw of \mathbf{Z}_j from its current posterior distribution (in the MCEM algorithm). For the Stochastic EM algorithm, there is only one Monte Carlo sample taken, so that $m = 1$ (in our MCEM notation) always.

Thus the Stochastic EM algorithm starts with arbitrary multinomial probabilities $\tau_{ij}^{(0)}$ ($i = 1, \dots, g$) for each observation j ; for convenience, initially these may be taken to be the

same for all j . Then with $\Psi^{(k)}$ denoting the value of Ψ obtained on the k th iteration, the following steps are used:

Stochastic E-step. For each j , a draw is made from the multinomial distribution with probabilities $\tau_{ij}^{(k)}$ ($i = 1, \dots, g$), which might be arbitrary for $k = 0$ as explained above and are given by the previous M-step for $k > 0$, as explained below. On the k th iteration, let the draw for the j th observation \mathbf{w}_j be $\mathbf{z}_j^{(1_k)} = (z_{1j}^{(1_k)}, \dots, z_{gj}^{(1_k)})^T$. Notice that exactly one of these $z_{ij}^{(1_k)}$ is 1 and the others 0 for each j . This results in a partition of the n observations $\mathbf{w}_1, \dots, \mathbf{w}_n$ with respect to the g components of the mixture.

M-step. Calculate

$$\pi_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^n z_{ij}^{(1_k)}, \quad (i = 1, \dots, g).$$

The calculation of $\theta^{(k+1)}$ depends upon the form of the densities f_i . The posterior probabilities required for the Stochastic E-step in the next iteration are updated as follows:

$$\begin{aligned} \tau_{ij}^{(k+1)} &= \tau_i(\mathbf{w}_j; \Psi^{(k+1)}) \\ &= \pi_i^{(k+1)} f_i(\mathbf{w}_j; \theta_i^{(k+1)}) / f(\mathbf{w}_j; \Psi^{(k+1)}) \quad (i = 1, \dots, g). \end{aligned}$$

Celeux and Diebolt (1985, 1986b) extend this to the case when g is not known and is to be estimated from the data. This algorithm prevents the sequence from staying near an unstable stationary point of the likelihood function. Thus it avoids the cases of slow convergence observed in some uses of the EM algorithm for the mixture problem. The sequence of estimates obtained by this procedure is an ergodic Markov chain and converges weakly to a stationary distribution, although the relation of this stationary distribution to the maxima of the likelihood is not known. Celeux and Diebolt (1990) propose a hybrid version of EM and simulated annealing, called the Simulated Annealing EM (SAEM), which resolves this difficulty. Lavielle and Moulines (1997) develop a similar technique wherein the convergence conditions are equivalent to those of EM. See Robert and Casella (2004) for more details.

Chauveau (1995) considers the asymptotic behavior of the stochastic EM algorithm in the case of a two-component normal mixture model fitted to censored data; see also Diebolt and Ip (1996). For an application of Stochastic EM to posterior simulation and model choice in Poisson panel count data, see Chib (1996).

6.4 DATA AUGMENTATION

6.4.1 The Algorithm

When the object of the Bayesian analysis is to estimate the entire posterior distribution and not just the posterior mode, Tanner and Wong (1987), Wei and Tanner (1990a, 1990b), and Tanner (1991, 1993) propose a method called the Data Augmentation algorithm. This is suitable when the incomplete-data posterior density is complicated, but the complete-data posterior density is relatively easy to handle and to draw from, just like the situation with likelihoods for which the EM algorithm is suitable.

Suppose that in the MCEM algorithm, the sequence $\{\Psi^{(k)}\}$ has converged to $\hat{\Psi}$. Then an estimate of the posterior distribution $p(\Psi | \mathbf{y})$ can be obtained as

$$\frac{1}{m} \sum_{j=1}^m p(\Psi | \mathbf{z}^{(j)}; \mathbf{y}),$$

where $\mathbf{z}^{(i)}$ denotes the i th draw from $p(\mathbf{z} | \mathbf{y}; \hat{\Psi})$. This is called the Poor Man's Data Augmentation algorithm 1 (PMDA-1). Thus the PMDA-1 is a noniterative algorithm and proceeds as follows:

1. Draw $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$ from $p(\mathbf{z} | \mathbf{y}; \hat{\Psi})$.
2. Compute an approximation to the posterior density as

$$\frac{1}{m} \sum_{j=1}^m p(\Psi | \mathbf{z}^{(j)}, \mathbf{y}). \quad (6.15)$$

This gives rise to a more general algorithm called the Data Augmentation algorithm. Here, starting from a prior density for the parameter Ψ , each stage of the algorithm constructs an estimate of the incomplete-data posterior density. Let this be $p^{(k)}(\Psi | \mathbf{y})$ at the end of the k th iteration. Choose and fix a positive integer m . Then the next iteration is as follows:

Imputation Step (I-step).

- (a1) Draw a $\Psi^{(j_k)}$ from the current approximation $p^{(k)}(\Psi | \mathbf{y})$.
 - (a2) Draw a $\mathbf{z}^{(j_k)}$ from the current approximation to the missing-data conditional density $p(\mathbf{z} | \mathbf{y}; \Psi^{(j_k)})$.
- Execute steps (a1) and (a2) for $j = 1, \dots, m$.

Posterior Step (P-step).

Set the next approximation to the posterior density by averaging the posterior densities obtained from the m draws of the missing data and using the supposed simpler form of the complete-data posterior density, as follows:

$$p^{(k+1)}(\Psi | \mathbf{y}) = \frac{1}{m} \sum_{j=1}^m p(\Psi | \mathbf{z}^{(j_k)}, \mathbf{y}).$$

Note that the I-step is equivalent to drawing $\mathbf{z}^{(1_k)}, \dots, \mathbf{z}^{(m_k)}$ from the predictive distribution $p(\mathbf{z} | \mathbf{y})$, by the method of composition using the identity

$$p(\mathbf{z} | \mathbf{y}) = \int_{\Omega} p(\mathbf{z} | \mathbf{y}; \Psi) p(\Psi | \mathbf{y}) d\Psi,$$

where Ω is the parameter space.

6.4.2 Example 6.5: Data Augmentation in the Multinomial (*Examples 1.1, 1.5 Continued*)

We return now to Example 1.1, which was developed as Example 1.5 in a Bayesian framework in Section 1.6.2. Suppose that the complete-data formulation is as before and that

the prior distribution for the parameter Ψ is beta (ν_1, ν_2). Then from (1.73), the posterior distribution for Ψ is

$$\text{beta}(\nu_1 + y_{12} + y_4, \nu_2 + y_2 + y_3). \quad (6.16)$$

Thus $p(\Psi | z^{(m_k)}, \mathbf{y})$ corresponds to (6.16) with y_{12} replaced by $y_{12}^{(m_k)}$, the m th draw of y_{12} on the k th iteration. The conditional distribution in (a2) above is binomial ($y_1, \Psi/(2+\Psi)$). If no knowledge of Ψ exists initially, one could start with a uniform prior which is a beta distribution with $\nu_1 = 1$ and $\nu_2 = 1$.

It is easily seen that the Data Augmentation algorithm is the iterative simulation version of the EM algorithm where the I-step is the analog of the E-step and the P-step is the analog of the M-step.

Suppose that in the Data Augmentation algorithm, we have $m = 1$. Then the algorithm can be considered to be a repeated application of the following two steps:

1. Given z^* , draw Ψ^* from $p(\Psi | z^*; \mathbf{y})$, in view of the relationship

$$p(\Psi | \mathbf{y}) = \int_{\mathcal{Z}} p(\Psi | z; \mathbf{y}) p(z | \mathbf{y}) dz.$$

2. Given Ψ^* , draw z^* from $p(z | \mathbf{y}; \Psi^*)$, in view of the relationship

$$p(z | \mathbf{y}) = \int_{\Omega} p(z | \mathbf{y}; \Psi) p(\Psi | \mathbf{y}) d\Psi.$$

This is called the Chained Data Augmentation algorithm.

Rubin (1991) discusses how techniques like multiple imputation, data augmentation, stochastic relaxation, and Sampling–Importance Resampling (SIR), combine simulation techniques with complete-data methods to attack problems that are difficult or impossible for EM, and illustrates them with the PET problem, which was introduced in Section 2.5.

In the context of the Data Augmentation algorithm, this technique of Importance Sampling could be used to improve the PMDA-1 algorithm if $p(z | \mathbf{y})$ is easy to evaluate. Note that equation (6.15) is only an approximation since the $z^{(i)}$ are sampled from $p(z | \mathbf{y}; \hat{\Psi})$ rather than from $p(z | \mathbf{y})$. In this context, Importance Sampling could be used to calculate the observed posterior as follows: Given $z^{(1)}, \dots, z^{(M)}$ from $p(z | \mathbf{y}; \hat{\Psi})$ assign weights

$$r_i = \frac{p(z^{(i)} | \mathbf{y})}{p(z^{(i)} | \mathbf{y}; \hat{\Psi})}$$

and replace equation (6.15) by a weighted average

$$\frac{\sum_{i=1}^m r_i p(\Psi | z^{(i)}; \mathbf{y})}{\sum_{i=1}^m r_i}.$$

Wei and Tanner (1990a) call this the Poor Man's Data Augmentation Algorithm 2 (PMDA-2). They also discuss approximations to this approach.

6.5 BAYESIAN EM

6.5.1 Posterior Mode by EM

Although we have focused on the application of the EM algorithm for computing MLEs in a frequentist framework, it can be equally well applied to find the mode of the posterior

distribution in a Bayesian framework, as discussed in Section 1.6.1. This problem is analogous to MLE and hence the EM algorithm and its variants can be adapted to compute MAP estimates.

If a prior $p(\Psi)$ is imposed on the parameter Ψ , then

$$\log L(\Psi) + \log p(\Psi)$$

is the log posterior function. Its maximum occurs at the posterior mode. The E-step is effectively the same as for the computation of the MLE of Ψ in a frequentist framework, requiring the calculation of the conditional expectation of the complete-data log likelihood (the Q -function). The M-step differs in that the objective function for the maximization process is equal to the Q -function, augmented by the log prior density. The combination of prior and sample information provides a posterior distribution of the parameter on which the estimation is based. The Q -function is now given by

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log L_c(\Psi) | \mathbf{y} \} + \log p(\Psi). \quad (6.17)$$

The difference

$$L(\Psi) + \log p(\Psi) - Q(\Psi; \Psi^{(k)}) = L(\Psi) - E_{\Psi^{(k)}} \{ \log L_c(\Psi) | \mathbf{y} \}$$

attains its maximum at $\Psi = \Psi^{(k)}$. Thus the M-step of maximizing $Q(\Psi; \Psi^{(k)})$ forces an increase in the log posterior function. The EM algorithm and its variants like the EM gradient algorithm are valuable in computing posterior modes, even more than in the computation of MLEs. These methods are useful in transmission tomography applications (Lange, 1999).

6.5.2 Example 6.6: Bayesian EM for Normal with Semi-Conjugate Prior

This example is from Gelman et al. (2004). Let us consider the problem of estimating a normal mean with unknown variance. Suppose we have a random sample $\mathbf{y} = (y_1, \dots, y_n)$ of n observations from $N(\mu, \sigma^2)$, so that $\Psi = (\mu, \sigma^2)^T$. Let the prior for μ be $N(\mu_0, \tau_0^2)$, with known parameters and for $\log \sigma$ be the standard noninformative uniform prior, leading to the prior on σ^2 to be $p(\sigma^2) \propto \sigma^{-2}$. This prior is not conjugate to the likelihood from the normal and is an example of a semi-conjugate prior. For this situation, there is no closed-form expression for the posterior marginal of μ ; nor is there a standard form for the posterior joint distribution of μ and σ^2 . However, using the EM algorithm, we can find the marginal posterior mode of μ averaging over σ^2 . For this, we need to maximize log posterior marginal density of μ with respect to μ .

The log posterior density (but for an additive constant not involving the parameters) is

$$\log p(\Psi | \mathbf{y}) = -\frac{1}{2} \left\{ \frac{(\mu - \mu_0)^2}{\tau_0^2} + (n+1) \log \sigma^2 + \frac{1}{\sigma^2} \sum_{j=1}^n (y_j - \mu)^2 \right\}. \quad (6.18)$$

E-Step. In the E-step we find the conditional expectation of (6.18) given data at the current value $\mu^{(k)}$ of μ . For this, we need to evaluate the expectations

$$E_{\mu^{(k)}}(\log \sigma), \text{ and } E_{\mu^{(k)}}\left(\frac{1}{\sigma^2}\right).$$

The first term does not depend on μ and will remain constant over various M-steps. The second term is evaluated by noting that

$$(1/\sigma^2) | \mu^{(k)}, \mathbf{y} \sim \chi_n^2(a_k),$$

where a_k is the noncentrality parameter given by

$$a_k^{-1} = \frac{1}{n} \sum_{j=1}^n (y_j - \mu^{(k)})^2,$$

and hence

$$E_{\mu^{(k)}}\left(\frac{1}{\sigma^2} \mid \mathbf{y}\right) = a_k.$$

Thus

$$E_{\mu^{(k)}}\{\log p(\Psi \mid \mathbf{y})\} = \text{constant} - \frac{1}{2} a_k \frac{(\mu - \mu_0)^2}{\tau_0^2} - \frac{1}{2} \sum_{j=1}^n (y_j - \mu)^2. \quad (6.19)$$

M-Step. The M-step is obtained by maximizing (6.19) with respect to μ . This is easily obtained by noting that it has the form of a log posterior density in the normal case with the prior $N(\mu_0, \tau_0^2)$ and variance $\frac{1}{n} \sum_{j=1}^n (y_j - \mu)^2$ with n observations y_j ($j = 1, \dots, n$).

This leads to the following iteration for the posterior mode

$$\mu^{(k+1)} = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^{2(k)}}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^{(k)2}}}, \quad (6.20)$$

where $\sigma^{(k)2} = \frac{1}{n} \sum_{j=1}^n (y_j - \mu^{(k)})^2$ and \bar{y} is the sample mean of y_1, \dots, y_n .

6.6 I.I.D. MONTE CARLO ALGORITHMS

6.6.1 Introduction

The preceding sections formulated various modifications of the standard EM algorithm in certain frequentist and Bayesian contexts. The methods suggested and the examples discussed involved drawing random samples from a variety of probability distributions. Algorithms for random sampling from some of these probability distributions like the normal and beta, are simple, direct, and well known. For others, more complicated and specialized methods may be needed. There are many reasons for this. We mention a few: (1) The distribution may not be one of the standard and well-known ones. (2) Because of the nature of the problem in which it arises, the distribution may be specified only up to a normalizing constant. For instance, quite often in Bayesian inference, the normalizing constants of the posterior densities involve complicated sums or integrals or their combinations, which are not analytically tractable. (3) In some Bayesian problems, even the form of the multivariate posterior densities may not be derivable analytically (let alone the normalizing constants) and thus they are specified only indirectly by certain conditional and marginal densities arising from the model and prior specifications. These specifications, however, should be adequate to result in a unique joint posterior density.

As is evident from the discussions in the book so far, the EM algorithm is typically employed for ML estimation, in contexts like missing data (or latent variables). In the Bayesian version of the same problems, the analog of the E-step would need to deal with the conditional densities of missing values or latent variables. This in turn results in the indirect

specification of the posterior joint distribution of model parameters and the missing values; such a specification would include several marginal and conditional distributions—the data distribution, the prior distribution, the missing value distribution, etc. As a consequence, often the posterior joint distribution becomes analytically quite intractable and computation of posterior mode, etc., become quite complex. Often, the only resort is a Monte Carlo method. These Monte Carlo exercises require special random sampling techniques.

In some of the simpler problems of this kind, the densities of the target posterior distributions are specified completely, although they may not be standard distributions. In such cases, often use is made of a surrogate probability distribution (often called a majorizing density function, an envelope, or a proposal distribution depending on the context) which is relatively easy to draw samples from, say, because it is a standard distribution. Then by a suitable acceptance-rejection mechanism, the selected samples are i.i.d. from the target distribution. We call such methods i.i.d. Monte Carlo methods (see Athreya, Delampady, and Krishnan, 2003). We now discuss some of these methods.

6.6.2 Rejection Sampling Methods

6.6.2.1 Rejection Sampling: Rejection Sampling was introduced by John von Neumann in 1951 (see Kennedy and Gentle, 1980). Lahiri (1951) introduced it in connection with probability-proportional-to-size (PPS) sampling from a finite population. Rejection Sampling is used when direct generation of a random sample from the target density is difficult or when the density is specified but for a constant as $f(w)$, but a related density $h(w)$ is available from which it is comparatively easy to generate random samples. This $h(w)$ is called the majorizing density function or an envelope or a proposal function and it should satisfy the condition that $Ch(w) \geq f(w)$ for every w , for some known constant $0 < C < \infty$. For the method to work, it should be easy to draw random samples from the distribution defined by the density function $h(\cdot)$.

Rejection Sampling Algorithm:

1. Draw w from $h(\cdot)$.
2. Draw u from uniform distribution on $(0, 1)$.
3. If $u \leq f(w)/\{Ch(w)\}$, then accept w as the desired realization; else return to Step 1. Repeat until one w is accepted.

Repeat this algorithm to select the desired number of samples. For distributions with finite support and bounded density, $h(\cdot)$ can always be chosen as uniform. If target f and proposal h functions are probability density functions, and $C(1 < C < \infty)$ is the value of a majorizing constant, then C can be interpreted as the average number of proposal variates required to obtain one sample unit from the target function. The probability of accepting a chosen proposal variate is large if the value of C is small. The requirements that must be satisfied in rejection sampling are:

1. the support of the target distribution must be a subset of the support of the proposal distribution;
2. the ratio of target and proposal functions should be bounded;
3. the constant C should be an upper bound to this ratio.

The success of random sample generation from a target function using the Rejection Sampling algorithm depends mainly on the choice of the proposal density function and the majorizing constant C .

6.6.2.2 Adaptive Rejection Sampling (ARS): When Rejection Sampling is intended to be used to generate random samples from a target function, finding an appropriate proposal density function and a majorizing constant C may not be easy. In the Adaptive Rejection Sampling method (Gilks, 1992; Gilks and Wild, 1992; Robert and Casella, 2004), the proposal density function is constructed as a (polygonal) envelope of the target function (on the log scale). The target function should be log-concave for this method to work. The envelope is updated whenever a proposal variate is rejected, so that the envelope moves closer to the target function by increasing the probability of accepting a candidate variate. The above references may be consulted for details of the ARS algorithm.

6.6.3 Importance Sampling

Importance Sampling (Geweke, 1989; Hesterberg, 1995) is another way to estimate the expectation of the function $E\{a(W)\}$ by drawing an independent sample w_1, \dots, w_n from a distribution of a given importance density $h(w)$, with $h(w) > 0$ and $a(w)f(w) \neq 0$ (whenever $f(w) > 0$). The integral can be estimated by

$$\tilde{E}\{a(W)\} = \frac{1}{n} \sum_{j=1}^n a(w_j)u(w_j),$$

where $u(w_j) = f(w_j)/h(w_j)$ ($j = 1, 2, \dots, n$) is a weight function (defined to be 0 when $f(w) = 0$), with estimated variance

$$\text{var}[\tilde{E}_f\{a(W)\}] = \frac{1}{n(n-1)} \sum_{j=1}^n \left[a(w_j)u(w_j) - \tilde{E}\{a(W)\} \right]^2.$$

The optimal importance density for minimizing the variance of the integration estimator is

$$h^*(w) = \frac{|a(w)|f(w)}{\int |a(v)|f(v)dv}.$$

The integration estimate can also be computed by the Importance Sampling ratio estimate

$$\hat{E}\{a(W)\} = \frac{\sum_{j=1}^n a(w_j)u(w_j)}{\sum_{j=1}^n u(w_j)},$$

and the corresponding variance estimate is

$$\text{var}(\hat{E}\{a(W)\}) = \frac{\sum_{j=1}^n [h(w_j) - \hat{E}\{a(W)\}]^2 \{u(w_j)\}^2}{\{\sum_{j=1}^n u(w_j)\}^2}.$$

The advantage of using the ratio estimate compared to the integration estimate is that in using the latter we need to know the weight function (that is, the ratio of target and importance functions) exactly, whereas in the former case, the ratio needs to be known only up to a multiplicative constant. If the support of the importance function consists of the support

of the density function $f(w)$, then the Importance Sampling estimator converges almost surely to the expectation.

A modified form of this algorithm, the Sampling–Importance Resampling (SIR) algorithm is useful in many contexts, one of which is drawing samples from an intractable distribution the form of which is known up to a multiplicative constant. This situation often arises when the constant in a posterior distribution needs complicated integration to be carried out. The SIR algorithm is related to the technique of rejection sampling. Let Ψ represent the parameter vector or parameter vector together with missing-data values as the case may be. In this technique, to draw samples from a distribution $f(\Psi)$, an approximation $h(\Psi)$ to it is chosen such that for a known constant C , $\{f(\Psi)/h(\Psi)\} \leq C$. Then the technique consists of the following steps:

1. Draw Ψ from $h(\Psi)$.
2. Draw u from the uniform distribution over $(0,1)$ or from any p.d.f. with appropriate modifications in Step 3 below.
3. If $u \leq \{f(\Psi)/Ch(\Psi)\}$, then accept Ψ ; otherwise go to Step 1.

Notice that the technique requires the calculation of the normalizing constants in f and h , as well as C .

For the SIR algorithm, numbers m and $M (> m)$ are chosen and fixed. A distribution $h(\Psi)$ is chosen which (a) should be easy to draw from; (b) is easy to evaluate up to a multiplicative constant; and (c) is an approximation to the distribution $f(\Psi)$ of Ψ . Then the algorithm consists of the following three steps:

1. Draw M values Ψ_1, \dots, Ψ_M from $h(\Psi)$.
2. Calculate the M ratios, called the importance ratios:

$$r(\Psi_j) = f(\Psi_j)/h(\Psi_j) \quad (j = 1, \dots, M).$$
3. Draw m values from the M values Ψ_1, \dots, Ψ_M in the above step with probability proportional to $r(\Psi_j)$.

It can be shown that, as $M/m \rightarrow \infty$, the probability distribution of the drawn values tends to the correct distribution. It is appropriate to use the SIR algorithm only if a reasonable approximation to the actual distribution is available. The advantage of the SIR algorithm over the rejection sampling method is that the normalizing constants of f and h and the constant C as in rejection sampling need not be evaluated. However, the draws in the SIR algorithm are only from the approximate distribution, whereas in rejection sampling they are from the actual distribution p .

To enhance the appeal of Monte Carlo integration methods, various techniques are used for making the approximation more accurate. One such technique is to choose $h(w)$ to be approximately proportional to $f(w)|a(w)|$. In Example 6.1 of Section 6.2.2, since $\cos \frac{\pi w}{2}$ is approximately $1 - \frac{\pi^2 w^2}{8}$ (two-term Taylor expansion) and $\frac{\pi^2}{8}$ being nearly one, a reasonable $h(w)$ is a probability density proportional to $(1 - w^2)$, that is, $\frac{3}{2}(1 - w^2)$. Results of such a simulation are given in Table 6.2 and Figure 6.2. This is Importance Sampling, the function $h(w)$ being called the Importance Function. Here sampling is made efficient by drawing from regions of higher density using the importance function. This is reflected in the variance being reduced by a factor of almost 100; see Lange (1999), Liu (2001), and Robert and Casella (2004) for interesting discussions of Monte Carlo integration, especially in statistical contexts.

Table 6.2 Descriptive Statistics of W with Density $\frac{3(1-w^2)}{2}$ and $V = \cos(\frac{\pi W}{2})$ Using Density of V Based on 10,000 Samples.

Statistic	W	V
No. of cases	10,000	10,000
Minimum	0.00011	0.48907
Maximum	0.99417	0.66667
Median	0.34258	0.64848
Mean	0.37245	0.63674
Standard Dev	0.24234	0.03180
Variance	0.05873	0.00101

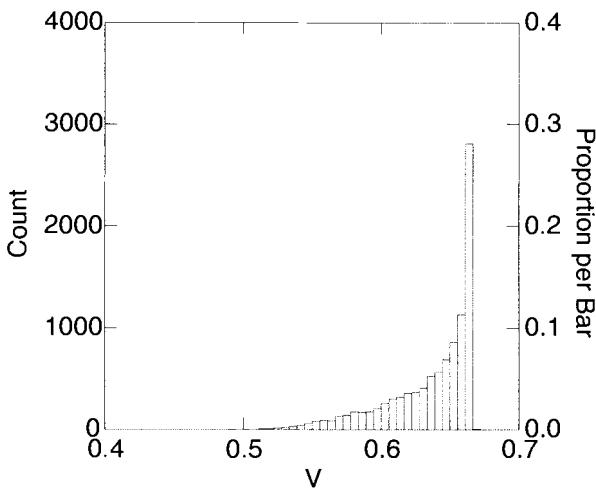


Figure 6.2 Histogram of $V = \cos(\frac{\pi W}{2})$ Using Density $\frac{3(1-w^2)}{2}$ Based on 10,000 Samples.

6.7 MARKOV CHAIN MONTE CARLO ALGORITHMS

6.7.1 Introduction

There are limitations to the applicability of Rejection Sampling and related algorithms. In practice, they are applicable only to generate from univariate distributions; the ARS algorithm is applicable only to generate from log-concave densities. Rejection sampling tends not to work well in high dimensions; the higher the dimension the closer the envelope or proposal needs to be to the target, for a reasonable acceptance proportion.

However, there is a substantial need to generate random samples from non-log-concave densities and from multivariate densities. In more complicated versions of problems needing Monte Carlo methods, especially in the multivariate case, the target distribution is often incompletely or indirectly specified, the specification depending on the way it arises in the

statistical problem on hand. However, so long as the target distribution is uniquely defined from the given specifications, it is possible to adopt an iterative random sampling procedure, which at the point of convergence delivers a random draw from the target distribution. Then the Monte Carlo methods needed are also consequently more complex. The techniques currently being used in such situations to draw samples from the target distributions involve generating realizations (sample paths) of a suitable ergodic Markov chain. The Markov chain and the method for generating realizations will depend on the specification of the target distribution. The Markov chain is so chosen that its limiting distribution is the target distribution. The generated realizations then yield random samples from the target distribution. Methods for constructing such Markov chains, methods for generating realizations from such Markov chains, and methods for using them to obtain random samples from the target distribution are called Markov chain Monte Carlo (MCMC) methods.

Our object in this chapter is to explore the EM-MCMC connections and to that end, we give a quick introduction to MCMC methods, leaving details and nuances to the reader to be figured out from the large number of references that we provide.

There are basically two schools of thought on how to deal with the dependence on starting values and to diagnose convergence. One school advocates generating a number of chains from different starting values in an attempt to aid exploration of the support of the stationary distribution and also to allow the use of between-chains convergence diagnostics. Another school advocates using just one very long chain with arguments including that this will be more tolerant of poor starting values and that MCMC convergence diagnostics are all rather fallible. A common compromise is to use three to five reasonably long chains, which allows the use of most diagnostics.

In some practical implementations of the MCMC methods, a burn-in period (b) and a gap (g) are used. Elements of the generated chain before the burn-in period are ignored, after which generated elements are picked up in gaps of g elements. Whatever be the method, there is a need to conduct diagnostic tests in an attempt to confirm that the Markov chains are converging to their stationary distributions and that the quantities being estimated are converging to their true values.

Geyer (1992) for instance, is an advocate of a single long chain; Gelman and Rubin (1992) for instance, are advocates of multiple short runs. An advantage of a single chain is that less burn-in is required, and it is less vulnerable to poor starting points; a disadvantage is that good convergence diagnostics are difficult to devise. Advantages of multiple chains are that convergence diagnostics are easier to devise, and parallel runs of several chains can be used to reduce run time; a disadvantage is that burn-ins throw away a large number of generated points.

Use of the generated Markov chain realizations is another issue. The purpose of the Monte Carlo exercise is generally the estimation of a few expectations like $E\{a(\Psi)\}$, the expectation of a function a of the parameters Ψ under some posterior distribution p . Towards this end, the precision of this estimation should be the focus when use is made of the MCMC runs. In a pair of articles, Gelman and Rubin (1992) and Geyer (1992) summarize many of the important issues in the implementation of MCMC methods. In particular, Gelman and Rubin (1992) note that the problem of creating a simulation mechanism is clearly separate from the problem of using this mechanism to draw inferences. They provide simple methods for obtaining inferences from iterative simulations, using multiple sequences that are generally applicable to the output of any iterative simulation. MacEachern and Berliner (1994) and Robert and Casella (2004) recommend the use of the entire generated chain for this purpose, although they form a dependent set. MacEachern and Berliner (1994) however, advocate subsampling of ‘independent’ units for purposes of convergence diagnostics.

Propp and Wilson (1996) introduce a method called **Exact Sampling or Perfect Sampling**, which uses a technique they call *Coupling from the Past (CFTP)*. This method is a modification of the MCMC method whereby the bias from the choice of the starting point of the chain is removed. This method is able to identify a point which could be regarded as a sample from the target distribution. However, it is slow in many cases.

Ripley (1987) provides a good general introduction to MCMC methods. An excellent source of work on MCMC methods is the trio of papers by Smith and Roberts (1993), Besag and Green (1993), and Gilks, Clayton, Spiegelhalter, Best, McNeil, Sharples, and Kirby (1993), which were read before the Royal Statistical Society at a meeting on Gibbs sampling and other MCMC methods. Tierney (1994) focuses on the theory of MCMC methods. Besag, Green, Higdon, and Mengersen (1995) provide an introduction to MCMC methods and its applications, along with several new ideas. Brooks and Roberts (1998) and Mengersen, Robert, and Guihenneuc-Joyaux (1999) give a review of convergence diagnostics, while a comprehensive account of MCMC methods is provided in the monograph on the topic by Gilks, Richardson, and Spiegelhalter (1996). A concise theoretical treatment of MCMC is provided in Gamerman and Lopes (2006). Marin and Robert (2007) deal with Bayesian computational issues; they provide code for Bayesian computations in the R language. Albert (2007) also provides R code.

6.7.2 Essence of MCMC

Here is the essence of the MCMC method. We follow the treatment in Athreya et al. (2003).

Given a probability distribution π on a set S , and a function h on S , suppose it is desired to compute the “integral of h with respect to π ”, which reduces to $\sum_j h(j)\pi_j$ in the countable case. One looks for an irreducible Markov chain $\{W_n\}$ with S as its state space and π as its stationary distribution. Then, starting from some initial value W_0 , run the Markov chain $\{W_j\}$ for a period of time, say $0, 1, 2, \dots, n$ and offer as an estimate

$$\mu_n = \frac{1}{n} \sum_0^{n-1} h(W_j). \quad (6.21)$$

By the Law of Large Numbers (LLN), which is valid for this type of Markov chain, this estimate μ_n will be close to $\sum_j h(j)\pi_j$ for large n . In particular, if one is interested in $\pi(B) \equiv \sum_{j \in B} \pi_j$ for some subset $B \subset S$ then by LLN this reduces to

$$\pi_n(B) \equiv \frac{1}{n} \sum_0^{n-1} I_B(W_j) \rightarrow \pi(B)$$

in probability as $n \rightarrow \infty$. In other words $\hat{\pi}_n(B)$ is the sample proportion of visits to B during $\{0, 1, 2, \dots, n-1\}$ by the Markov chain.

If an irreducible Markov chain $\{W_n\}$ with a countable state space S is also *aperiodic*, then, in addition to the LLN, the following result on the convergence of $\text{pr}(W_n = j)$ holds, namely, that

$$\sum_j |\text{pr}(W_n = j) - \pi_j| \rightarrow 0 \quad (6.22)$$

as $n \rightarrow \infty$, for any initial distribution of W_0 . This means that for large n the probability distribution of W_n is close to π .

There is a result similar to (6.22) for the general state space case that asserts that under suitable conditions, the probability distribution of W_n is close to π as $n \rightarrow \infty$.

This suggests that instead of doing one run of length n , one could do N independent runs each of length m so that $n \cong Nm$ and then from the j th run use only the m th observation, say, $W_{m,j}$ and offer the estimate

$$\tilde{\mu}_{N,m} \equiv \frac{1}{N} \sum_{j=1}^N h(W_{m,j}). \quad (6.23)$$

There are other variations as well.

We now discuss particular types of MCMC algorithms that are frequently used in statistical analysis.

6.7.3 Metropolis–Hastings Algorithms

The roots of MCMC methods can be traced back to an algorithm of Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), popularly known as the Metropolis algorithm; see also Metropolis and Ulam (1949). Its initial use was in physics and chemistry to investigate the equilibrium properties of large systems of particles, such as molecules in a gas. The first use of the Metropolis algorithm in a statistical context was by Hastings (1970), who developed it further. Besag (1974) studied the associated Markov field structure. It took a long time before it came to be exploited in Bayesian inference by Gelfand and Smith (1990).

Let S be a finite or countable set. Let π be a probability distribution on S . (π is called the target distribution.) Let $A \equiv ((a_{ij}))$ be a transition probability matrix such that for each i , it is computationally easy to generate a sample from the distribution $\{a_{ij} : j \in S\}$. Then generate a Markov chain $\{U_n\}$ as follows.

Step 1: If $U_n = i$, first sample from the distribution $\{a_{ij} : j \in S\}$ and denote that observation V_n .

Step 2: Choose U_{n+1} from the two values U_n and V_n according to the probability distribution

$$\text{pr}(U_{n+1} = V_n | U_n, V_n) = \rho(U_n, V_n),$$

$$\text{pr}(U_{n+1} = U_n | U_n, V_n) = 1 - \rho(U_n, V_n), \quad (6.24)$$

where the “acceptance probability” $\rho(\cdot, \cdot)$ is given by

$$\rho(i, j) = \min \left\{ \frac{\pi_j}{\pi_i} \frac{a_{ji}}{a_{ij}}, 1 \right\} \quad (6.25)$$

for all (i, j) such that $\pi_i a_{ij} > 0$. It is not difficult to verify that $\{U_n\}$ is a Markov chain with transition probability matrix $P = ((p_{ij}))$ given by

$$p_{ij} = \begin{cases} a_{ij} \rho(i, j) & j \neq i \\ 1 - \sum_{k \neq i} p_{ik}, & j = i. \end{cases}$$

The probability a_{ij} is called the “proposal transition probability” and $\rho(i, j)$ the “acceptance probability”. A most useful feature of this transition mechanism P is that P and π satisfy the so called *detailed balance* condition:

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \text{for all } i, j. \quad (6.26)$$

This implies that for any j

$$\sum_i \pi_i p_{ij} = \pi_j \sum_i p_{ji} = \pi_j. \quad (6.27)$$

That is, π is a stationary probability distribution for P .

Now assume that S is irreducible with respect to A and $\pi_i > 0$ for all i in S . Then it can be shown that P is irreducible and since it has a stationary distribution π and LLN is available. This algorithm is thus a very flexible and useful one. The choice of A is subject only to the condition that S is irreducible with respect to A . Clearly, it is no loss of generality to assume that $\pi_i > 0$ for all i in S . A sufficient condition for the aperiodicity of P is that $p_{ii} > 0$ for some i or equivalently

$$\sum_{j \neq i} a_{ij} \rho(i, j) < 1.$$

A sufficient condition for this is that there exists a pair (i, j) such that $\pi_i a_{ij} > 0$ and $\pi_j a_{ji} < \pi_i a_{ij}$.

Recall that if P is aperiodic then both the LLN and (6.22) hold. If S is not finite or countable but is a continuum and the target distribution $\pi(\cdot)$ has a density $p(\cdot)$ then one proceeds as follows: Let A be a transition function such that for each u , $A(u, \cdot)$ has a density $a(u, v)$. Then proceed as in the discrete case but set the “acceptance probability” $\rho(u, v)$ to be

$$\rho(u, v) = \min \left\{ \frac{p(v)a(v, u)}{p(u)a(u, v)}, 1 \right\}$$

for all (u, v) such that $p(u)a(u, v) > 0$. Another useful feature of the above algorithm is that it is enough to know $\{\pi_i\}$ up to a multiplicative constant as in the definition of “the acceptance probability” $\rho(\cdot, \cdot)$, only the ratios $\frac{\pi_i}{\pi_j}$ need to be calculated. This is useful in (i) Bayesian statistical applications of MCMC for calculating the moments of the posterior distribution of the parameters given the data; and (ii) in *image processing* and statistical mechanics where the set S consists of configurations over a multidimensional grid of pixels where in each pixel there is a fixed number of levels and the probability distribution is specified via a Gibbs potential function whose normalizing constant (“partition function”) is not easy to compute.

Chib and Greenberg (1995) provide a tutorial on the Metropolis-Hastings algorithm; see also Gilks et al. (1998) and Liu (2001).

There are several approaches to selecting proposal functions, resulting in specific types of M-H algorithms. Two of these types are: Random Walk Metropolis–Hastings algorithm and Independent Metropolis–Hastings algorithm, which we describe below.

6.7.3.1 Independent Metropolis–Hastings Algorithm Under a proposal distribution $a(v | u)$ using an independence chain, the probability of moving to a point v is independent of the current position u of the chain, that is, $a(v | u) = a(v)$. The acceptance probability can be written as $\alpha(u, v) = \min\{\frac{w(v)}{w(u)}, 1\}$, where $w(u) = \frac{\pi(u)}{a(u)}$ can

be considered a weight function that can be used in an importance sampling process, if the generated variates are from $a(\mathbf{v} \mid \mathbf{u})$. It is suggested that the proposal be selected in such a way that the weight function is bounded. In that case, the generated Markov chain is uniformly ergodic. It must be ensured that the support of the proposal distribution covers the support of the target distribution.

6.7.3.2 Random Walk Metropolis–Hastings Algorithm Under a proposal distribution $a(\mathbf{v} \mid \mathbf{u})$ using a random walk chain, the new value \mathbf{v} equals the current value plus ϵ_k . A random variate ϵ_k follows the distribution $a(\cdot)$ and is independent of the current value, that is, $a(\mathbf{v} \mid \mathbf{u}) = a(\mathbf{v} - \mathbf{u})$. The acceptance probability can be written as $\alpha(\mathbf{u}, \mathbf{v}) = \min\{\frac{\pi(\mathbf{v})}{\pi(\mathbf{u})}, 1\}$, where the generated variates are from the proposal distribution. When the proposal density is continuous and positive around zero, the generated RWM–H chain is ergodic. It is recommended that a symmetric proposal distribution around zero be used. The performance of the algorithm depends on the scale of the proposal distribution. A proposal with small steps has a high acceptance rate, but the generated chain mixes very slowly. A proposal with large steps does not move frequently, resulting in a low acceptance rate and slow mixing. The scale of the proposal should be chosen so that both of the above cases are avoided. If the range of the target function is finite, then boundary points can be treated as reflecting barriers of a random walk in the algorithm.

6.8 GIBBS SAMPLING

6.8.1 Introduction

Gibbs sampling (Casella and George, 1992) is a special case of the Metropolis–Hastings algorithm. It deals with the problem of random sampling from a multivariate distribution, which is defined in terms of a collection of conditional distributions of its component random variables or subvectors, in such a way that this collection uniquely defines the target joint distribution. Gibbs sampling can be regarded as component-wise Metropolis–Hastings algorithm. The formulation of defining marginal and conditional distributions often arises naturally in Bayesian problems in terms of the observable variable distributions and the prior distributions. These then give rise to the posterior distribution of the parameters. This posterior distribution is often difficult to work out analytically, necessitating the development of Monte Carlo procedures like Gibbs sampling. One of the simpler situations, called the case of full conditionals, is where the defining collection consists of the conditional distribution of each single (univariate or multivariate) component of the random vector given the values of the rest.

A version of the Gibbs sampling algorithm suitable for the case of full conditionals for generating from a multivariate distribution of $\mathbf{W} = (\mathbf{W}_1, \dots, \mathbf{W}_d)^T$ can be formulated as follows; here \mathbf{W}_i ($i = 1, \dots, d$) may itself be multidimensional. In applications like Bayesian computations, some of the components of this vector will be parameters and some will be missing observations, prior distributions, etc.

The algorithm is as follows:

Starting from an initial value $\mathbf{W}^{(0)} = (\mathbf{W}_2^{(0)}, \dots, \mathbf{W}_d^{(0)})^T$, carry out the following d steps on the k th iteration:

We use π as a generic symbol for the distribution of the variables involved.

- (1) Draw $\mathbf{W}_1^{(k+1)}$ from $\pi(\mathbf{W}_1 \mid \mathbf{W}_2^{(k)}, \dots, \mathbf{W}_d^{(k)})$.

- (2) Draw $\mathbf{W}_2^{(k+1)}$ from $\pi(\mathbf{W}_2 | \mathbf{W}_1^{(k+1)}, \mathbf{W}_3^{(k)}, \dots, \mathbf{W}_d^{(k)})$.
 \vdots
(d) Draw $\mathbf{W}_d^{(k+1)}$ from $\pi(\mathbf{W}_d | \mathbf{W}_1^{(k+1)}, \dots, \mathbf{W}_{d-1}^{(k+1)})$.

The above Gibbs sampling procedure, starting from $\mathbf{w}_2^{(0)}, \dots, \mathbf{w}_d^{(0)}$, generates what is called a ‘Gibbs sequence’ $\mathbf{w}_1^{(1)}, \dots, \mathbf{w}_d^{(1)}, \dots, \mathbf{w}_1^{(n)}, \dots, \mathbf{w}_d^{(n)}, \dots$. This sequence is a realization of a homogeneous Markov chain with a stationary distribution, which is the unique multivariate distribution defined by the full conditionals. Thus for large n , $\mathbf{w}_1^{(n)}, \dots, \mathbf{w}_d^{(n)}$ can be considered as a random draw from the target multivariate distribution.

Many interesting properties of such a Markov sequence have been established, including geometric convergence as $n \rightarrow \infty$ to a distribution $\pi(\mathbf{W}_1^{(n)}, \dots, \mathbf{W}_d^{(n)})$, which is the stationary distribution of the Markov chain; see Roberts and Polson (1994).

There are other variations of the Gibbs sampling algorithm which may involve draws from some marginal distributions as well. Anyhow, the collection of marginal and conditional distributions used for the draws should ensure unique existence of the joint distribution coinciding with the target distribution.

The Gibbs sampling method is also useful to approximate the marginal density $f(\mathbf{w}_i)$ ($i = 1, \dots, d$) of a joint density function $f(\mathbf{w}_1, \dots, \mathbf{w}_d)$ or its parameters by averaging the final conditional densities of each sequence—for that matter, marginal multivariate densities and their parameters.

As discussed in Besag and Green (1993), Gibbs sampling was founded on the pioneering ideas of Grenander (1983). It was not until the following year that Gibbs sampling was so termed by Geman and Geman (1984) in their seminal paper on image processing. The Metropolis-Hastings algorithm seems to have had little use in statistics until the advent of Gibbs sampling by Geman and Geman (1984). Even then, Gibbs sampling, which is a special case of the Metropolis-Hastings algorithm, seems to have been used mainly in applications in spatial statistics (for example, as in Besag, York, and Mollié, 1991), until the appearance of the paper by Gelfand and Smith (1990). They brought into focus its tremendous potential in a wide variety of statistical problems. In particular, they observed that almost any Bayesian computation could be carried out via Gibbs sampling. Later, Geyer and Thompson (1992) make a similar point about ML calculations; almost any ML computation can be done by some MCMC scheme. Further applications of Gibbs sampling methods have been explored in the papers by Gelfand, Hills, Racine-Poon, and Smith (1990) (normal data modeling); Zeger and Karim (1991) (general linear random effects models); Gelfand, Smith, and Lee (1992) (truncated data and constrained parameters); Gelfand and Carlin (1993) (constrained and missing data); Gilks et al. (1993) (complex modeling in medicine); Sobel and Lange (1994) (pedigree analysis); and Carlin and Chib (1995) (Bayesian model choice), among several others.

The reader is referred to Casella and George (1992) for a tutorial on Gibbs sampling, to Arnold (1993) for an elementary introduction to it, and to Ritter and Tanner (1992) for methodology on facilitating Gibbs sampling.

6.8.2 Rao–Blackwellized Estimates with Gibbs Samples

Let us consider a statistical problem of drawing inferences about a parameter Ψ of the probability density function $f(\mathbf{w}|\Psi)$, based on a random sample $\mathbf{w}_1, \dots, \mathbf{w}_n$. By the sufficiency principle, a statistic T is sufficient for Ψ if the conditional distribution of the

sample given the statistic does not involve the parameter Ψ . By the Rao–Blackwell theorem, the conditional expectation of the estimator given a sufficient statistic is an improved estimator; that is, when $\delta(\mathbf{w}_1, \dots, \mathbf{w}_n)$ is an estimator of Ψ with finite variance,

$$\text{var}[E\{\delta(\mathbf{w}_1, \dots, \mathbf{w}_n) \mid \mathbf{T} = \mathbf{t}\}] \leq \text{var}\{\delta(\mathbf{w}_1, \dots, \mathbf{w}_n)\}.$$

Using a conditional expectation with respect to another estimator as an improved estimator is often called Rao–Blackwellization, even if the conditioning statistic is not a sufficient statistic. This leads to the use of the Rao–Blackwellized estimator

$$\delta_{\text{RB}} = \frac{1}{n} \sum_{i=1}^n E\{h(\mathbf{w}_{1j}^{(k)} \mid \mathbf{w}_{2j}, \dots, \mathbf{w}_{pj})\}$$

instead of the empirical estimator

$$\frac{1}{n} \sum_{i=j}^n \{h(w_{1j}^{(k)})\}$$

in the Gibbs sampling method above; see Liu, Wong, and Kong (1994) and Robert and Casella (2004) for details.

As is evident, the M–H algorithm stated above is applicable for drawing from a multidimensional distribution as well. A variation of the M–H algorithm in that case is the single component M–H algorithm, where the updating is done coordinate-wise, keeping the other coordinates at their last values. This means that updating is done by means of conditional distributions of each coordinate given the rest. Gibbs sampling can be regarded as a variation of this theme, where the target distribution is itself defined in terms of a full set of conditional distributions from which it is easy to draw samples, and which uniquely specify the target joint distribution.

6.8.3 Example 6.7: Why Does Gibbs Sampling Work?

This is a toy example to show why Gibbs Sampling works; Gibbs sampling is too inefficient a procedure for drawing bivariate normal samples; very efficient direct methods are available. However, for purposes of illustration, we shall discuss now how to draw random samples from the bivariate normal using Gibbs Sampling. We shall simplify the problem by considering

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

It is a property of the bivariate normal that

$$(U \mid V = v) \sim N(\rho v, 1 - \rho^2),$$

$$(V \mid U = u) \sim N(\rho u, 1 - \rho^2).$$

Using this property, Gibbs sampling proceeds as follows:

1. Start from an arbitrary value u_0 for U .
- Repeat the following steps for $i = 0, 1, \dots, n$.
- Given u_i for U , draw a random sample from $N(\rho u_i, 1 - \rho^2)$ and denote it v_i .

3. Given v_i for V , draw a random sample from $N(\rho v_i, 1 - \rho^2)$ and denote it u_{i+1} .

Thus we have $(u_i, v_i), i = 0, 1, 2, \dots, n$. The theory of Gibbs sampling tells us that if n is large, then (u_n, v_n) can be considered to be a random sample from

$$N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

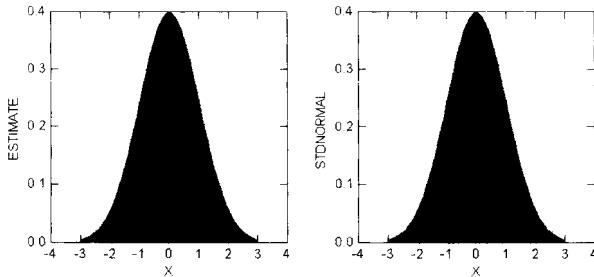


Figure 6.3 Histograms of Gibbs Sampled (left) and Actual (right) $N(0, 1)$.

Why does Gibbs sampling work? A necessary condition for it to work is that the marginals and conditionals that are used in the scheme should imply the targeted joint distribution uniquely. This is satisfied in this case and we are not proving that here. Let us explain why Gibbs sampling works in this example. Let us consider the transition from U_i to U_{i+1} . This takes place through an intermediate value of v_i . The transition function from $U_i = u_i$ to V_i is given by $\phi(v_i; \rho u_i, 1 - \rho^2)$ and the transition function from $V_i = v_i$ to U_{i+1} is given by $\phi(u_{i+1}; \rho v_i, 1 - \rho^2)$. Thus the transition function from $U_i = u_i$ to U_{i+1} is given by the convolution of these integrals as

$$\int_{-\infty}^{\infty} \phi(v_i; \rho u_i, 1 - \rho^2) \phi(u_{i+1}; \rho v_i, 1 - \rho^2) dv_i.$$

With some effort, this integral can be simplified to be

$$p(u_i, u_{i+1}) = \phi(u_{i+1}; \rho^2 u_i, (1 - \rho^4)).$$

Thus it is clear that the sequence $\{U_i\}$ is a homogeneous Markov chain with $p(u_i, u_{i+1})$ as the transition function, since the form of this transition function does not depend on i .

Let us explain this. Suppose we pretend that we have generated u_0 from the correct marginal. Then since we have used the correct conditional, v_0 is bound to be a sample from the correct marginal. This is just by construction. By the same argument, u_1 has the correct marginal, and so on. Further, it can be shown that

$$\phi(v; 0, 1) = \int_{-\infty}^{\infty} p(u, v) \phi(u; 0, 1) dx,$$

showing that the invariant distribution of the Markov chain is $N(0, 1)$. Hence it does not also matter where we start from; u_0 need not be from the (unknown) marginal! Thus Gibbs

sampling when it has converged produces a random sample from the marginal distribution of U .

We could estimate such quantities as the marginal density $f(u)$ of U or the mean μ_u of U using this sample. However, since we know the form of the conditional distribution $f(u \mid v)$, we can use the estimator

$$\hat{f}(u) = \frac{1}{n} \sum_{j=1}^n f(u \mid v_j),$$

which in our case becomes

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \phi(x; \rho v_i, 1 - \rho^2).$$

These conditional density function values carry more information on $f(u)$ than the values u_1, \dots, u_n themselves in view of the Rao–Blackwell theorem. Similarly, if the marginal mean μ_u of U is to be estimated, then

$$\hat{\mu}_u = \frac{1}{n} \sum_{j=1}^n E(U \mid v_j) = \frac{\rho}{n} \sum_{j=1}^n v_i$$

is better (has less variance) than $\frac{1}{n} \sum_{j=1}^n u_j$.

We use this Gibbs sampling to generate a random sample of 1000 from

$$N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right),$$

using $n = 5000$, that is using the 5000th value in the sample Markov chain path as one sample unit. We repeated this independently 1000 times to generate the 1000 sample units. In Figure 6.3, we present the density of x from these 1000 values and the actual density of $N(0, 1)$ representing the marginal distribution of U , respectively. We notice that the histogram is a very good approximation to the standard normal curve.

6.9 EXAMPLES OF MCMC ALGORITHMS

6.9.1 Example 6.8: M-H Algorithm for Bayesian Probit Regression

This example is from Chib (2004). The investigation is about the effect of three factors (covariates) on the proportion of mothers infected during Caesarean births. The response variable y_j is a binary variable with $y_j = 1$ if the j th mother is infected, and is 0 otherwise. The covariates are: x_1 : Caesarean planned (1) or not (0); x_2 : risk factors present (1) or not (0) at time of birth; x_3 : antibiotics given (1) or not (0). A Bayesian probit regression exercise is conducted. The data used for the analysis appear in Table 6.3.

The model used and the prior distribution for the Bayesian analysis are as follows. Let $\mathbf{x}_j = (1, x_{1j}, x_{2j}, x_{3j})^T$. The probability of infection is modeled as: $\text{pr}(Y_j = 1 \mid \mathbf{x}_j; \boldsymbol{\beta}) = \Phi(\mathbf{x}_j^T \boldsymbol{\beta})$; $\boldsymbol{\beta} \sim N_4(\mathbf{0}, 10\mathbf{I}_4)$. Let us assume that the y_j are conditionally independent. Denoting the data by \mathbf{y} , the posterior can be written as

$$\pi(\boldsymbol{\beta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\beta}) \prod_{j=1}^{251} \{\Phi(\mathbf{x}_j^T \boldsymbol{\beta})\}^{y_j} \{1 - \Phi(\mathbf{x}_j^T \boldsymbol{\beta})\}^{(1-y_j)}.$$

Table 6.3 Infection During Caesarean Births with Covariates.

x_1	x_2	x_3	No. infected	No. of deliveries
1	1	1	11	98
0	1	1	1	18
0	0	1	0	2
1	1	0	23	26
0	1	0	28	58
1	0	0	0	9
0	0	0	8	40

where $\pi(\beta)$ is $N(\mathbf{0}, 10\mathbf{I}_4)$ density.

The MLE's of the parameters are obtained in order to serve as starting values for the M-H algorithm as suggested by Gelman and Rubin (1992) as follows:

Parameter	β_0	β_1	β_2	β_3
Estimate	-1.0930	0.6076	1.1975	-1.9047

We use the following random walk proposal distribution,

$$\beta^{(t)} \mid \beta^{(t-1)} \sim N_4(\beta^{(t-1)}, \mathbf{V}),$$

where \mathbf{V} can be chosen to be

$$\begin{pmatrix} 0.040745 & -0.007038 & -0.039399 & 0.004829 \\ & 0.073102 & -0.006940 & -0.050162 \\ & & 0.062292 & -0.016803 \\ & & & 0.080788 \end{pmatrix}.$$

The acceptance probability for M-H is:

$$\min \left\{ \frac{\pi(\beta^{(t)})}{\pi(\beta^{(t-1)})}, 1 \right\}.$$

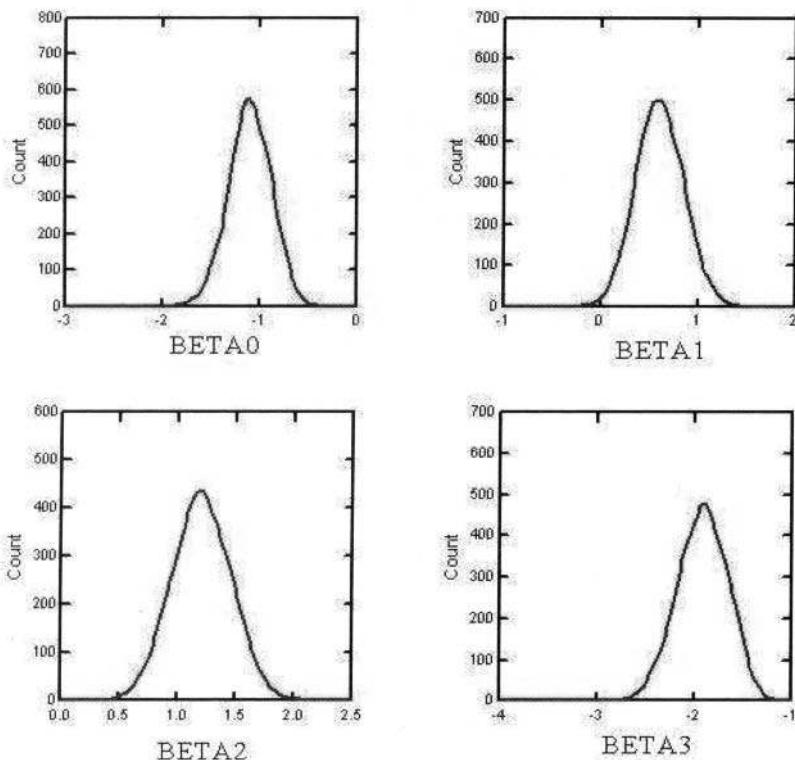
With this the following are summary results based on the generated samples. The burn-in and gap were chosen heuristically. Monte Carlo errors of the regression estimates were computed from the 5000 draws of the respective β 's. No diagnostics were actually attempted. The purpose here is only to illustrate an instance of the M-H algorithm.

6.9.2 Monte Carlo EM with MCMC

Sometimes in the E- and/or M-step, the distributions involved may not be amenable to direct sampling, but may be complex enough to necessitate indirect sampling by MCMC algorithms. Here is an example of Monte Carlo EM with MCMC in the M step.

Table 6.4 Summary from 5,000 Metropolis Draws; Burn-in: 1,000; Gap: 21.

	Prior		Posterior					
	Mean	St Dev	Mean	St Dev	2.5% Point	Median	97.5% Point	MC error
β_0	0	3.162	-1.106	0.219	-1.548	-1.103	-0.691	0.0031
β_1	0	3.162	0.609	0.245	0.142	0.601	1.103	0.0035
β_2	0	3.162	1.208	0.255	0.714	1.203	1.707	0.0036
β_3	0	3.162	-1.912	0.259	-2.433	-1.906	-1.430	0.0037

**Figure 6.4** Kernel Densities of M-H generated $\beta_0, \beta_1, \beta_2, \beta_3$ posterior.

Maximum likelihood estimation in a Generalized Linear Mixed Model (GLMM) can be handled by the EM algorithm, if necessary by Monte Carlo E- and M-steps. A typical way of invoking the EM algorithm is to assume the random effects to be missing data, whence the problem considerably simplifies in view of the random effects becoming fixed upon filling them in the E-step; see Section 5.9.1. We give an example from McCulloch and Searle (2001), where the Monte Carlo M-step is implemented by a Metropolis-Hastings algorithm.

The model is

$$y_j \mid \mathbf{u} \sim \text{indep } f_j(y_j \mid \mathbf{u}) \quad (j = 1, \dots, n);$$

that is, the Y_j are conditionally independent given the u_j , where

$$\begin{aligned} f_j(y_j \mid \mathbf{u}_j) &= \exp[\kappa^{-1}\{\theta_j y_j - b(\theta_j)\} + c(y_j; \kappa)], \\ E(y_j \mid \mathbf{u}) &= \mu_j, \\ h(\mu_j) &= \mathbf{x}_j^T \boldsymbol{\beta} + \mathbf{z}_j^T \mathbf{u}_j, \\ \mathbf{u} &\sim f(\mathbf{u} \mid \mathbf{D}), \end{aligned}$$

and \mathbf{D} represents the parameters in the distribution of \mathbf{U} . Note that \mathbf{z}_j is a design vector and so is not denoting any missing data as in most instances in the rest of the book.

The complete-data vector \mathbf{x} is given by $(\mathbf{y}^T, \mathbf{u}^T)^T$, where $\mathbf{y} = (y_1, \dots, y_n)^T$ and $\mathbf{u} = (u_1, \dots, u_n)^T$. The distribution of the complete data can be written, using $f(\mathbf{y}, \mathbf{u}) = f(\mathbf{u})f(\mathbf{y} \mid \mathbf{u})$, so that the complete-data log likelihood is

$$\begin{aligned} \log L_c(\boldsymbol{\Psi}) &= \log f(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}) + \log f(\mathbf{u}; \mathbf{D}) = \sum_{j=1}^n \log f(y_j \mid \mathbf{u}; \boldsymbol{\beta}) + \log f(\mathbf{u}; \mathbf{D}) \\ &= \kappa^{-1}\left\{\sum_{j=1}^n \theta_j y_j - b(\theta_j)\right\} + \sum_{j=1}^n \log c(y_j; \kappa) + \log f(\mathbf{u} \mid \mathbf{D}). \end{aligned} \quad (6.28)$$

This is because conditional on \mathbf{u} , the y_j are independent. Then $\boldsymbol{\beta}$ and τ appear in the first part (the GLM part) and \mathbf{D} appears in the second part. The maximization can then be done separately, the first part imitating the fixed effects GLM and the second part similar to maximum likelihood with respect to $f(\mathbf{u}; \mathbf{D})$, which can be simple if f is a member of the exponential family; see Section 5.10.6. Thus an iteration of the EM algorithm reduces to

1. finding $\boldsymbol{\beta}^{(k+1)}$ and $\kappa^{(k+1)}$ to maximize $E_{\boldsymbol{\Psi}^{(k)}}\{\log f(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \kappa) \mid \mathbf{y}\}$,
2. finding $\mathbf{D}^{(k+1)}$ to maximize $E_{\boldsymbol{\Psi}^{(k)}}\{\log f(\mathbf{u} \mid \mathbf{D}^{(k)}) \mid \mathbf{y}\}$.

The advantage of this procedure is the avoidance of the calculation of the likelihood (f_Y) and making do with the conditional distribution. Even so, it may not always be easy to derive analytically expressions for the two expectations involved. In such cases, we may resort to Monte Carlo methods both for estimating these integrals (expectations) and for maximizing. Thus the algorithm consists in drawing a random sample $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(M)}$ from the conditional distribution and updating the parameters $\boldsymbol{\beta}$, κ , \mathbf{D} by

1. calculating the updated $\boldsymbol{\beta}^{(k+1)}$ and $\kappa^{(k+1)}$ to maximize the Monte Carlo estimate

$$\frac{1}{M} \sum_{m=1}^M \log f(\mathbf{y} \mid \mathbf{u}^{(m)}; \boldsymbol{\beta}, \kappa)$$

of

$$E_{\boldsymbol{\Psi}^{(k)}}\{\log f(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \kappa) \mid \mathbf{y}\},$$

and

2. calculating the updated $\mathbf{D}^{(k+1)}$ to maximize

$$\frac{1}{m} \sum_{j=1}^m \log f(\mathbf{u}^{(m)} \mid \mathbf{D}).$$

The Metropolis-Hastings algorithm is used to sample from the conditional distribution of \mathbf{U} . In the Metropolis-Hastings algorithm, if we choose the proposal to be the marginal density $f(\mathbf{u})$ of \mathbf{U} , then the M-H probability of selection turns out to be

$$\begin{aligned} \frac{f(\mathbf{u}^* | \mathbf{y}, \boldsymbol{\beta}, \kappa, \mathbf{D}) f(\mathbf{u})}{f(\mathbf{u} | \mathbf{y}, \boldsymbol{\beta}, \kappa, \mathbf{D}) f(\mathbf{u}^*)} &= \frac{\prod_{j=1}^n f(y_j | \mathbf{u}^*, \boldsymbol{\beta}, \kappa) f(\mathbf{u}^* | \mathbf{D}) f(\mathbf{u} | \mathbf{D})}{\prod_{j=1}^n f(y_j | \mathbf{u}, \boldsymbol{\beta}, \kappa) f(\mathbf{u} | \mathbf{D}) f(\mathbf{u}^* | \mathbf{D})} \\ &= \frac{\prod_{j=1}^n f(y_j | \mathbf{u}^*, \boldsymbol{\beta}, \kappa)}{\prod_{j=1}^n f(y_j | \mathbf{u}, \boldsymbol{\beta}, \kappa)}. \end{aligned} \quad (6.29)$$

This computation only involves the conditional distribution of \mathbf{Y} given \mathbf{u} . There are a number of ways of using MCMC in these contexts. McCulloch (1994, 1997) uses Gibbs sampling for probit models and the Metropolis-Hastings algorithm for GLMM and Booth and Hobert (1999) use an independent sampler.

In MCEM there is a problem of quantifying the Monte Carlo error in the E-step. Although the MCEM circumvents evaluating a complicated E-step, it necessitates a difficult choice of the sample size m at each MCE step. If m is too small, the Monte Carlo error will be large; a large m results in a waste of resources. Ideally, m is chosen by evaluating the asymptotic variance of the Monte Carlo estimator of the expectation in the MCE step and on the applicability of the Central Limit Theorem for this estimator. In problems involving high-dimensional intractable integrals in the Q -function, MCMC methods are often invoked; in these situations invoking the Central Limit Theorem and evaluating asymptotic variances are difficult problems. There are modifications to these methods to save computing resources—for instance, one may gradually increase m over iterations. Booth and Hobert (1999), McCulloch (1997), and Levine and Casella (1998) discuss these issues.

6.9.3 Example 6.9: Gibbs Sampling for the Mixture Problem

Gibbs sampling is extensively used in many Bayesian problems where the joint distribution is complicated and is difficult to handle, but the conditional distributions are often easy enough to draw from. We give an example here.

Consider a Bayesian version of the finite mixture problem such as the one discussed in Section 2.7. If $\boldsymbol{\theta}_i$ denotes the parameter vector of the component density $f_i(\mathbf{w}; \boldsymbol{\theta}_i)$ in a mixture of the form

$$f(\mathbf{w}; \boldsymbol{\Psi}) = \sum_{i=1}^g \pi_i f_i(\mathbf{w}; \boldsymbol{\theta}_i),$$

then the totality of parameters for the problem is

$$\boldsymbol{\Psi} = (\boldsymbol{\theta}^T, \boldsymbol{\pi}^T)^T,$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \dots, \boldsymbol{\theta}_g^T)^T$ and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{g-1})^T$. If the component densities have common parameters, then $\boldsymbol{\theta}$ is the vector of those parameters known *a priori* to be distinct. Given an observed random sample $\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T$ from the mixture distribution and a prior density $p(\boldsymbol{\Psi})$ for $\boldsymbol{\Psi}$, the posterior density for $\boldsymbol{\Psi}$,

$$p(\boldsymbol{\Psi} | \mathbf{y}) \propto \prod_{j=1}^n \left\{ \sum_{i=1}^g \pi_i f_i(\mathbf{w}_j; \boldsymbol{\theta}_i) \right\} p(\boldsymbol{\Psi}),$$

is not easy to handle (Titterington et al., 1985). However, suppose we formulate a complete-data problem as in Section 2.7 by introducing the missing-data vector $\mathbf{z} = (z_1^T, \dots, z_n^T)^T$,

where \mathbf{z}_j is the vector containing the g zero-one indicator variables that define the component membership of each \mathbf{w}_j . Then the components of the conditional model in a Bayesian framework are

$$p(\mathbf{w}_j \mid \mathbf{z}_j, \Psi) = \prod_{i=1}^g p_i^{z_{ij}}(\mathbf{w}_j; \boldsymbol{\theta}_i) p(\mathbf{z}_j \mid \boldsymbol{\pi})$$

where

$$p(\mathbf{z}_j \mid \boldsymbol{\pi}) = \prod_{i=1}^g \pi_i^{z_{ij}}.$$

If $p(\Psi)$ is taken to be a product of a Dirichlet prior for $\boldsymbol{\pi}$ and an independent prior for $\boldsymbol{\theta}$, then the conditional distributions of $(\boldsymbol{\pi} \mid \boldsymbol{\theta}, \mathbf{z})$, $(\boldsymbol{\theta} \mid \boldsymbol{\pi}, \mathbf{z})$, and $(\mathbf{z} \mid \boldsymbol{\theta}, \boldsymbol{\pi})$ are all easily defined and handled so that a Gibbs sampling procedure can be used to obtain estimates of an otherwise intractable posterior distribution; see Gelman and King (1990), Smith and Roberts (1993), Diebolt and Robert (1994), Escobar and West (1995), and Robert (1996) for more details.

6.9.4 Example 6.10: Bayesian Probit Analysis with Data Augmentation

This example is from Albert and Chib (1993). The data consist of $n = 50$ cases where a Bernoulli variable Y_j is observed together with observations $\mathbf{x}_j = (x_{1j}, x_{2j}, x_{3j})^T$, ($j = 1, \dots, 50$) on three predictor variables. It is assumed that the Y_j are independently distributed according to a Bernoulli distribution,

$$Y_j \sim \text{Bernoulli}(p_j) \quad (j = 1, \dots, n),$$

where the p_j 's are allowed to depend on \mathbf{x}_j through a linear function $\mathbf{x}_j^T \boldsymbol{\beta}$. It is assumed further that z_1, \dots, z_n are unobservable random variables, distributed independently according to a normal distribution with mean $\mathbf{x}_j^T \boldsymbol{\beta}$ and unit variance, where $y_j = 1$ if $z_j > 0$ and $y_j = 0$ if $z_j < 0$; that is,

$$p_j = 1 - \Phi(\mathbf{x}_j^T \boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is an unknown vector of parameters. The complete-data vector \mathbf{x} is $(\mathbf{y}^T, \mathbf{z}^T)^T$, where $\mathbf{y} = (y_1, \dots, y_n)^T$ contains the observed data and $\mathbf{z} = (z_1, \dots, z_n)^T$ is the vector of missing observations.

Then with $\pi(\boldsymbol{\beta})$ as prior for $\boldsymbol{\beta}$, the posterior is

$$\pi(\boldsymbol{\beta}, \mathbf{Z} \mid \mathbf{y}) \propto \pi(\boldsymbol{\beta}) \prod_{j=1}^n \{I_{(0, \infty)}(z_j)\}^{y_j} \{I_{(-\infty, 0)}(z_j)\}^{1-y_j} \phi(Z_i; \mathbf{x}_i^T \boldsymbol{\beta}, 1).$$

Let us use the following normal linear model for regression

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X} = (x_1^T, \dots, x_n^T)^T$; $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \mathbf{I}_n)$.

If we use the **diffuse prior** of $\boldsymbol{\beta}$, the posterior is:

$$\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{Z} \sim N_k(\hat{\boldsymbol{\beta}}_{\mathbf{Z}}, (\mathbf{X}^T \mathbf{X})^{-1}),$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{Z}} = (\mathbf{X}^T \mathbf{X})^{-1}(\mathbf{X}^T \mathbf{Z})$, and

$$Z_i \mid \mathbf{y}, \boldsymbol{\beta} \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, 1), \text{ truncated at the left by 0 if } y_i = 1;$$

$Z_i | \mathbf{y}, \boldsymbol{\beta} \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, 1)$, truncated at the right by 0 if $y_i = 0$.

For a proper conjugate prior $N(\boldsymbol{\beta}^*, \mathbf{B}^*)$, the posterior is:

$$\boldsymbol{\beta} | \mathbf{y}, \mathbf{Z} \sim N_k(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{B}}),$$

where

$$\tilde{\boldsymbol{\beta}} = (\mathbf{B}^{*-1} + \mathbf{X}^T \mathbf{X})^{-1} (\mathbf{B}^{*-1} \boldsymbol{\beta}^* + \mathbf{X}^T \mathbf{Z}).$$

$$\tilde{\mathbf{B}} = (\mathbf{B}^{*-1} + \mathbf{X}^T \mathbf{X})^{-1}.$$

The computations for obtaining posterior mean vector estimate $\hat{\boldsymbol{\beta}}$ for diffuse prior, are implemented through Gibbs sampling as follows:

The data file used for this example contains $n = 50$ cases of x_1, x_2, x_3, y . The initial value of $\boldsymbol{\beta}$ is generated from least-squares linear regression of y on x_1, x_2, x_3 with a constant. But the β_0 value is ignored. The initial values are taken to be: $\beta_1^0 = 0.18; \beta_2^0 = 0.11; \beta_3^0 = 0.13$.

Initial Computation:

Form a 50×3 matrix \mathbf{X} with i th row as x_{i1}, x_{i2}, x_{i3} .

Compute $\mathbf{X}^T \mathbf{X}$ and $(\mathbf{X}^T \mathbf{X})^{-1}$.

Gibbs Sampling:

Step 1. For each $i, i = 1, \dots, 50$:

Generate $W_i \sim N(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, 1)$.

For $y_i = 1$, reject W_i if $W_i \leq 0$; generate again; if $W_i > 0$, then set $Z_i = W_i$.

For $y_i = 0$, reject W_i if $W_i > 0$; generate again; if $W_i \leq 0$, then set $Z_i = W_i$.

Form a column vector $\mathbf{Z} = (Z_1, \dots, Z_{50})^T$.

Step 2. Now we have $x_{i1}, x_{i2}, x_{i3}, Z_i, i = 1, \dots, 50$.

Compute $\hat{\boldsymbol{\beta}}_Z = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{Z})$.

Generate $\boldsymbol{\beta} \sim N_3(\hat{\boldsymbol{\beta}}_Z, (\mathbf{X}^T \mathbf{X})^{-1})$.

With this updated $\boldsymbol{\beta}$ go to Step 1.

Repeat the process until convergence.

Data used for these computations are given in Table 6.5. The results of this Gibbs sampling are given in Table 6.6 and Figure 6.5.

6.9.5 Example 6.11: Gibbs Sampling for Censored Normal

Suppose we have a random sample of n from $N(\mu, \sigma^2)$, with the first m observed (w_1, \dots, w_m), and the remaining censored at c (that is, we only know they are $> c$ and not the actual values). Let us assume a uniform (noninformative) prior for μ and σ^2 . Let us denote the observations by \mathbf{y} . The likelihood function and the posterior are given by

$$L(\mu, \sigma^2 | \mathbf{y}) \propto \frac{\prod_{j=1}^m \phi(w_j | \mu, \sigma^2)}{\left\{ \Phi\left(\frac{c-\mu}{\sigma}\right) \right\}^{n-m}} \left\{ 1 - \Phi\left(\frac{c-\mu}{\sigma}\right) \right\}^{n-m}. \quad (6.30)$$

Table 6.5 Data Used for Gibbs Sampling for Bayesian Probit Analysis with Data Augmentation.

0.279	-0.270	-0.206	1	-1.372	0.202	-0.394	1	0.331	0.481	0.263	1
-0.765	-0.321	-0.177	0	0.452	0.322	-0.643	0	-1.261	-0.214	0.424	0
-0.724	0.333	-1.105	0	0.364	0.735	0.095	0	-0.269	-0.666	0.354	1
0.423	0.656	0.362	0	-0.741	0.412	-0.950	0	-1.818	0.187	-0.480	0
0.033	-1.362	1.267	0	-0.143	0.280	0.013	0	0.864	-0.315	-1.194	0
-0.770	1.635	-0.146	0	-1.759	-1.420	0.376	1	-0.494	0.242	0.430	1
-0.018	0.518	-0.360	1	-0.145	-0.388	0.408	0	-0.043	1.575	-0.504	1
0.251	0.566	-0.612	1	1.383	0.704	0.580	1	-0.005	0.511	-0.416	1
-1.425	-0.397	-0.286	0	-0.169	0.213	-0.620	0	-2.382	-0.245	-0.452	0
-1.284	-3.019	-0.357	0	-2.002	-1.288	0.241	0	-1.005	-0.560	-1.450	0
-0.178	0.404	1.343	1	1.375	0.918	1.097	0	-1.634	-1.031	-1.461	1
-1.044	0.230	-0.409	0	0.502	0.892	-0.010	0	0.780	1.468	-0.010	0
1.478	-0.044	1.794	1	0.732	0.339	-2.030	0	0.344	1.290	-0.614	1
0.715	1.197	0.649	1	0.017	1.707	0.106	0	-0.466	-0.431	-0.320	0
-0.689	-0.334	0.088	1	0.021	-0.760	-0.242	1	0.257	-0.151	-0.204	0
1.103	0.084	1.697	1	-0.601	0.529	1.100	0	0.926	-1.153	-1.353	0
0.443	0.236	-0.242	0	-1.578	-1.321	-2.448	0				

Table 6.6 Summary from Gibbs Draws for Posterior with Sample Size: 5,000; Burn-in: 100; Gap: 21.

Statistic	Direct Estimates			RB Estimates		
	β_1	β_2	β_3	β_1	β_2	β_3
N	5000	5000	5000	5000	5000	5000
Mean	0.217	-0.026	0.427	0.217	-0.027	0.429
St dev	0.227	0.230	0.151	0.144	0.155	0.173
St error	0.003	0.003	0.003	0.002	0.002	0.002
2.5% point	-0.237	-0.466	-0.026	-0.070	-0.303	0.124
Median	0.216	-0.027	0.421	0.214	-0.029	0.416
97.5% point	0.658	0.424	0.936	0.527	0.269	0.802

This is analytically difficult to maximize and even to sample from. Let us augment the data with missing data z , the values of the censored observations. The complete-data vector is $\mathbf{x} = (w_1, \dots, w_m, w_{m+1}, w_n)^T$ is given by $(\mathbf{y}^T, \mathbf{z}^T)^T$, where $\mathbf{y} = (w_1, \dots, w_m)^T$ and $\mathbf{z} = (w_{m+1}, \dots, w_n)^T$. The complete-data log likelihood function can be expressed as

$$\log L_c(\mathbf{x} | \mu, \sigma^2) = \sum_{j=1}^m \log \phi(w_j; \mu, \sigma^2) + \sum_{i=m+1}^n \log \phi(w_i; \mu, \sigma^2).$$

Now the full conditionals for Gibbs sampling can be written as:

1. The conditional density of \mathbf{z} given μ, σ^2 , and \mathbf{y} is given by the product of the truncated normal densities,

$$\frac{\phi(w_j; \mu, \sigma^2)}{1 - \Phi\left(\frac{c-\mu}{\sigma}\right)}.$$

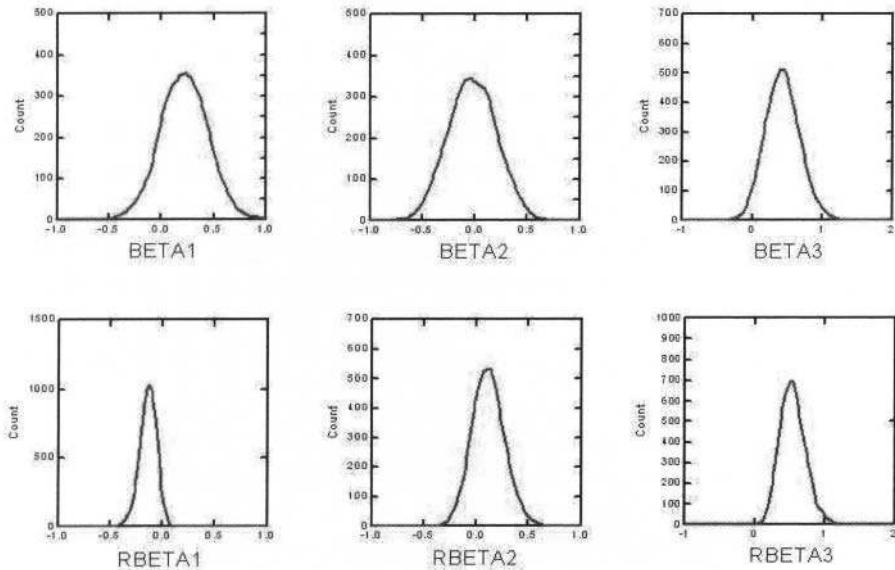


Figure 6.5 Kernel Densities of Gibbs Generated $\beta_1, \beta_2, \beta_3$; Direct Estimates in Top Row; Rao-Blackwellized (RB) Estimates in Bottom Row.

2. The conditional posterior distribution of σ^2 given μ, z , and y is obtained through the product

$$\prod_{j=1}^n \phi(w_j; \mu, \sigma^2),$$

which is proportional to

$$\sigma^{-n} \exp \left\{ -\frac{\sum_{j=1}^n (w_j - \mu)^2}{2\sigma^2} \right\},$$

the scaled inverted χ^2 distribution with scale parameter $\sum_{j=1}^n (w_j - \mu)^2$ and $(n - 2)/2$ as shape parameter;

3. The conditional distribution of μ given σ^2, z , and y is obtained through the product

$$\begin{aligned} \prod_{j=1}^n \phi(w_j | \mu, \sigma^2) &\propto \exp \left\{ -\frac{\sum_{j=1}^n (w_j - \mu)^2}{2\sigma^2} \right\} \\ &= \exp \left\{ -\frac{\sum_{j=1}^n \{w_j - \bar{x}\} + \bar{x} - \mu\}^2}{2\sigma^2} \right\}, \end{aligned}$$

and

$$(\mu | \sigma^2, \mathbf{z}, \mathbf{y}) \sim N(\bar{w}, \frac{\sigma^2}{n}).$$

4. Rao-Blackwellization yields good estimates.

Numerical Example

In the following data set, $m = 43$, $n = 50$, $a = 15$:

9.97	7.42	10.46	9.06	1.30	11.24	4.05	2.54	0.74	5.15	11.10
4.22	3.27	12.30	10.56	8.28	7.41	7.43	8.15	9.48	7.25	5.69
11.88	8.44	12.61	13.54	13.49	1.48	14.52	8.25	11.17	8.44	12.73
1.44	10.51	6.17	11.43	6.05	6.72	4.72	8.76	8.88	5.20	

We ran a Gibbs sampler on this data set and generated 100,000 random samples from the posterior of μ , σ . We used a burn-in period of 1000 and a gap of 100. We present some descriptive statistics and histograms of the generated variables in Table 6.7 and Figure 6.6.

Table 6.7 Descriptive Statistics of Gibbs-Sampling-Generated Posterior μ and σ .

	μ	σ
No. of cases	1,000,000	1,000,000
Minimum	5.737	2.982
Maximum	13.770	9.508
Range	8.033	6.526
Median	9.306	4.866
Mean	9.312	4.920
95% CI Upper	9.314	4.922
95% CI Lower	9.311	4.919
Standard Dev	0.712	0.583
Variance	0.508	0.340
C.V.	0.077	0.118
Skewness(G1)	0.060	0.598
Kurtosis(G2)	0.194	0.703

6.10 RELATIONSHIP OF EM TO GIBBS SAMPLING

6.10.1 EM–Gibbs Sampling Connection

We discuss in this section the strong connection between the EM algorithm and Gibbs sampling as formulated by Robert and Casella (2004). Just as the EM algorithm is a statistically tuned method of ML estimation in incomplete-data problems in a frequentist framework, Gibbs sampling is a method for solving the Bayesian analogs of such problems. The EM algorithm can be regarded as a precursor to Gibbs sampling in missing data models (Andrieu, Doucet, and Robert, 2004). There is a similarity in the two-stage procedures in both these algorithms. Both exploit the knowledge and the possible simplicity of the

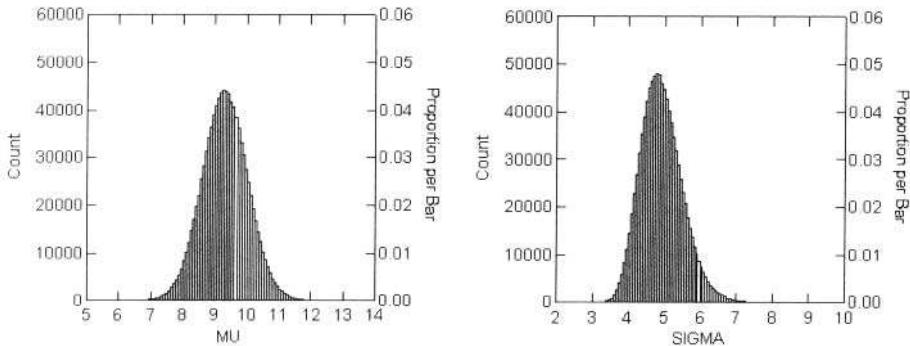


Figure 6.6 Histograms of posterior μ and σ using Gibbs Sampling.

conditional distribution of the missing variables. The connection goes further, as explained below.

Let us use our notation for the general formulation of the EM algorithm, where $L_c(\Psi)$ and $L(\Psi)$ denote the complete-data and incomplete-data likelihood functions formed from x and y , respectively, and z denotes the missing data. We define

$$L^*(\Psi | y, z) = \frac{L_c(\Psi)}{\int L_c(\Psi) d\Psi} \quad (6.31)$$

and

$$L^*(\Psi | y) = \frac{L(\Psi)}{\int L(\Psi) d\Psi} \quad (6.32)$$

assuming the integrals in the denominators of (6.31) and (6.32) to be finite; that is, the complete- and incomplete-data likelihoods can be normalized. Now create the two-stage Gibbs sampling scheme:

$$z | \Psi \sim k(z|y; \Psi),$$

$$\Psi | z \sim L^*(\Psi | y, z),$$

where

$$k(z|y; \Psi) = L_c(\Psi)/L(\Psi).$$

This is a direct connection to EM based on L_c and k . The E-step is replaced by generating a random sample from k . The M-step is replaced by generating from L^* .

The validity of the EM-Gibbs connection can be established, following Robert and Casella (2004), as follows: Let the transition kernel of the Markov chain generated by Gibbs sampling be

$$\Upsilon(\Psi, \Psi' | y) = \int_z k(z | y; \Psi) L^*(\Psi') dz.$$

It can be shown (Robert and Casella, 2004; Problem 9.15 on p. 363) that the invariant distribution of the Markov chain is the incomplete-data likelihood. That is,

$$L(\Psi') = \int_{\Psi} \Upsilon(\Psi, \Psi' | y) L(\Psi) d\Psi.$$

Hence if $L_c(\Psi)$ is integrable in Ψ , so is $L(\Psi')$, and hence the invariant distribution is a proper density. So the Markov chain is positive, and convergence follows from the theorem that under the positivity condition, if the transition kernel

$$K\{(x, y), (x', y')\} = f_{X|Y}(x' | y)f_{Y|X}(y' | x'),$$

is absolutely continuous with respect to the dominating measure, the chain is Harris recurrent and ergodic with stationary distribution f .

Whenever EM works in the frequentist version, data augmentation works in the Bayesian version of the problem for sampling from the posterior density; for example, gene frequency estimation problem. (See Lange, 1999.)

Both can be justified by using Markov chain theory. The incomplete-data likelihood is a solution to the integral equation of successive substitution sampling; Gibbs sampling can then be used to calculate the likelihood function.

It can be shown that $L^*(\Psi | \mathbf{y})$ is the solution of

$$L^*\Psi | \mathbf{y}) = \int \left\{ \int L^*(\Psi | \mathbf{y}, \mathbf{z})k(\mathbf{z} | \Psi', \mathbf{y})d\mathbf{z} \right\} L^*(\Psi' | \mathbf{y})d\Psi'.$$

The sequence $\Psi^{(k)}$ from the Gibbs iteration

$$\Psi^{(k)} \sim L^*(\Psi | \mathbf{y}, \mathbf{z}^{(k-1)}),$$

$$\mathbf{z}^{(k)} \sim k(\mathbf{z} | \mathbf{y}; \Psi^{(k)}),$$

converges to a random variable with density $L^*(\Psi | \mathbf{y})$ as $k \rightarrow \infty$. This can be used to compute the likelihood function $L(\Psi)$.

Based on the same functions $L(\Psi)$ and $k(\mathbf{z} | \Psi, \mathbf{y})$, the EM algorithm will get the MLE from $L(\Psi)$, whereas Gibbs sampling will get us the entire function.

See Casella and Berger (1994) and Smith and Roberts (1993) for this likelihood implementation. Baum and Petrie (1966) and Baum et al. (1970) make the likelihood-Markov chain connection quite apparent.

6.10.2 Example 6.12: EM–Gibbs Connection for Censored Data from Normal (*Example 6.11 Continued*)

In this example (Robert and Casella, 2004), the density of a missing observation is

$$\frac{\phi(z - \mu)}{1 - \Phi(c - \mu)}$$

and the distribution of $\mu | \mathbf{y}, \mathbf{z}$ is given by

$$L(\mu | \mathbf{y}, \mathbf{z}) \propto \prod_{j=1}^m e^{-(w_j - \mu)^2/2} \prod_{j=m+1}^n e^{-(w_j - \mu)^2/2},$$

which corresponds to a

$$N\left(\bar{w}, \frac{1}{n}\right)$$

distribution. This shows that L^* exists and that we can run Gibbs sampling.

6.10.3 Example 6.13: EM–Gibbs Connection for Normal Mixtures

Here we explore the close connection between the EM algorithm and Gibbs sampling with an example. The example is the mixture of two univariate normals, where for ML estimation of parameters the EM algorithm is often used. In the EM formulation, the group labels z_{ij} are the latent or missing values, where $z_{ij} = 1$ if y_j comes from the i th group and is zero otherwise ($i = 1, 2; j = 1, \dots, n$). We put $\mathbf{z}_j = (z_{1j}, z_{2j})^T$ and $\mathbf{z} = (\mathbf{z}_1^T, \dots, \mathbf{z}_n^T)^T$. The Bayesian analog would consider unknown parameters as well as the missing values as the random variables of the posterior distribution. Thus Gibbs sampling would involve sampling from the posterior joint density of $(\mu_1, \mu_2, \mathbf{Z})$, where \mathbf{Z} is the random vector corresponding to \mathbf{z} . Let us assume flat independent priors for the parameters (μ_1, μ_2) . Let us consider the case of a mixture of two univariate normals with known variances σ_1^2 and σ_2^2 and known mixing proportions π_1 and π_2 with $\pi_1 + \pi_2 = 1$, with unknown means μ_1 and μ_2 . Let $\mathbf{y} = (y_1, \dots, y_n)^T$ contain the observations. Then the Gibbs sampling procedure is:

1. Start with some initial values $\mu_1^{(0)}$ and $\mu_2^{(0)}$ for μ_1 and μ_2 .
2. At iteration $(k+1)$,
 - (a) since the conditional distribution of Z_{1j} given μ_1 and μ_2 is Bernoulli with probability parameter

$$\tau_{1j}^{(k)} = \frac{\pi_1 \phi(y_j; \mu_1^{(k)}, \sigma_1^2)}{\pi_1 \phi(y_j; \mu_1^{(k)}, \sigma_1^2) + \pi_2 \phi(y_j; \mu_2^{(k)}, \sigma_2^2)},$$

generate $z_{1j}^{(k)} = 1 - z_{2j}^{(k)}$ from this distribution;

- (b) since the conditional distribution of μ_1 and μ_2 are normal let

$$\nu_i^{(k+1)} = \frac{\sum_{j=1}^n z_{ij}^{(k)} y_j}{\sum_{j=1}^n z_{ij}^{(k)}} \quad (i = 1, 2),$$

and draw $\mu_1^{(k+1)} \sim N(\nu_1^{(k+1)}, \sigma_1^2)$ and $\mu_2^{(k+1)} \sim N(\nu_2^{(k+1)}, \sigma_2^2)$.

3. Continue Step 2 increasing k by 1, until convergence by a suitable criterion.

Steps 2(a) and (b) are the same as the E- and M-steps except that we are now drawing random samples rather than computing expectations or maximizing.

6.10.4 Rate of Convergence of Gibbs Sampling and EM

We follow the treatment in Sahu and Roberts (1999). In the case of Gibbs sampling, convergence is that of a sequence of distribution iterates to the stationary target distribution. Let $\Upsilon(\Psi, \Psi')$ denote the transition kernel density associated with a Gibbs sampling Markov chain and $\pi(\Psi)$ be the unique stationary density of the chain. Let Υ^k denote the transition density at the k th iteration. Let $\|\cdot\|$ denote a suitable divergence measure between densities, such as total variation distance, χ^2 divergence, and Kullback-Leibler divergence; let $V(\cdot)$ be a suitable function of the parameter vector Ψ . Then the rate of convergence ρ of Gibbs sampling is

$$\rho = \min_{\xi \in [0, 1]} \|\Upsilon^k(\Psi^{(0)}, \Psi) - \pi(\Psi)\| \leq V(\Psi^{(0)})\xi^{(k)}. \quad (6.33)$$

For a nonstochastic algorithm like the EM, the rate of convergence can be defined by

$$\rho = \min_{\xi \in [0,1]} \| \Psi^{(k+1)} - \Psi^{(k)} \| \leq V(\Psi^{(0)}) \xi^{(k)}. \quad (6.34)$$

The role of ρ in both these cases is similar, although the algorithms are different in nature. The idea of this definition is that by inverting (6.33) or (6.34) as the case may be, an assessment can be made of the number of iterations needed for convergence corresponding to a requirement that $\| \Psi^{(k+1)} - \Psi^{(k)} \| < \epsilon$, for a given ϵ as

$$(\log \rho)^{-1} \log(\epsilon/V(\Psi^{(0)})).$$

The convergence properties of these two algorithms are similar. We have seen earlier how the EM algorithm is similar in spirit to a two-stage Gibbs sampling. An extension of the EM algorithm is an ECM algorithm (Expectation–Conditional Maximization) (Meng and Rubin, 1993; Sexton and Swensen, 2000), where the maximization step is implemented in several steps, each step updating some of the parameters by conditional maximizations (Section 5.2). Thus an ECM algorithm can be regarded as analogous to multi-step Gibbs sampling. Working in a Gaussian setup, Sahu and Roberts (1999) prove that the rates of convergence of ECM and corresponding Gibbs sampling are the same. The implication of this result is that convergence properties and the number of iterations needed in one can be used in the other. One of their theorems is as follows: Assuming that the target distribution of both EM (ECM) and Gibbs sampling $\pi(\Psi, Z)$ is a multivariate normal distribution, the rates of convergence of EM (ECM) and Gibbs sampling are the same.

There are other connections between convergence rates of the EM-type algorithms and Gibbs sampling. For instance, there are similarities in the structures of the Jacobian matrices associated with the ECM algorithm and Gibbs sampling. As we have seen in Chapter 3, these Jacobian matrices determine the rates of convergence of the algorithms and Amit (1991), in fact, computes the rates of convergence of Gibbs sampling using the Jacobian matrix; see Meng and Rubin (1992); Meng (1994).

6.11 DATA AUGMENTATION AND GIBBS SAMPLING

6.11.1 Introduction

Suppose in a Bayesian problem we are interested in the posterior distribution $p(\Psi | \mathbf{y})$ of parameter Ψ and because of its analytical intractability, we wish to approximate it by MCMC. It is sometimes the case that the full conditional posterior $p(\Psi_i | \Psi_{-i}, \mathbf{y})$ (where Ψ_{-i} denotes all parameters except the i th component of Ψ), is not available in a closed form or otherwise not of a form easy to sample from. In such contexts the technique of data augmentation (Tanner and Wong, 1987) in the spirit of the EM algorithm is often useful. The idea is to suitably augment the data with ‘missing’ data \mathbf{z} to make the conditional $p(\Psi_i | \Psi_{-i}, \mathbf{z}, \mathbf{y})$ easy to sample from and by Gibbs sampling draw inference from the posterior

$$p(\Psi | \mathbf{y}) = \int p(\Psi | \mathbf{z}, \mathbf{y}) p(\mathbf{z} | \mathbf{y}) d\mathbf{z}.$$

The following example is adapted from Sørensen and Gianola (2002).

6.11.2 Example 6.14: Data Augmentation and Gibbs Sampling for Censored Normal (Example 6.12 Continued)

Continuing with Example 6.12, and assuming a uniform (noninformative) prior distributions for μ and σ^2 , the posterior distribution of μ and σ^2 is given by the likelihood function (6.30), but for a multiplicative constant not involving the parameters. This posterior distribution is not analytically easy to handle for integration towards investigation of either joint moments or even marginal moments, nor even for sampling using such devices as Gibbs sampling. However, upon augmenting the data with missing values $\mathbf{z} = (w_{m+1}, \dots, w_n)^T$, the actual values of the censored observations, a complete-data vector becomes $\mathbf{x} = (w_1, \dots, w_m, w_{m+1}, \dots, w_n)^T = (\mathbf{y}^T, \mathbf{z}^T)^T$, which is easy to handle, since it consists of i.i.d. $N(\mu, \sigma^2)$. The complete-data log likelihood is then, but for an additive constant not involving the parameters,

$$\log L_c(\mu, \sigma^2) = \sum_{j=1}^m \log \phi(w_j; \mu, \sigma^2) + \sum_{j=m+1}^n \log \phi(w_j; \mu, \sigma^2). \quad (6.35)$$

Under the assumption of a uniform prior, (6.35) is also the logarithm of the joint posterior of μ, σ^2 , and \mathbf{z} , except for an additive constant.

Now it is possible to work out the full conditionals for Gibbs sampling for μ, σ^2 , and \mathbf{z} . The conditional posterior distribution for $\mathbf{z} = (w_{m+1}, \dots, w_n)^T$ given μ, σ^2 , and \mathbf{y} is the product of the truncated densities,

$$\prod_{j=m+1}^n \frac{\phi(w_j; \mu, \sigma^2)}{1 - \Phi\left(\frac{c-\mu}{\sigma}\right)}. \quad (6.36)$$

The conditional posterior distribution of σ^2 given μ, \mathbf{z} , and \mathbf{y} is (but for a multiplicative constant not involving the parameters),

$$\prod_{j=1}^n \phi(w_j | \mu, \sigma^2) \propto \sigma^{-n} \exp\left\{-\frac{\sum_{j=1}^n (w_j - \mu)^2}{2\sigma^2}\right\}, \quad (6.37)$$

which makes this distribution a scaled inverted gamma distribution with scale parameter $\sum_{j=1}^n (w_j - \mu)^2$ and shape parameter $(n - 2)/2$. The conditional distribution of μ given σ^2, \mathbf{z} , and \mathbf{y} is obtained through the product,

$$\prod_{j=1}^n \phi(w_j | \mu, \sigma^2),$$

which is proportional to

$$\exp\left[-\sum_{j=1}^n \frac{(w_j - \mu)^2}{2\sigma^2}\right] = \exp\left[-\sum_{j=1}^n \frac{\{(w_j - \bar{w}) + \bar{w} - \mu\}^2}{2\sigma^2}\right].$$

This gives the required conditional distribution,

$$\mu | \sigma^2, \mathbf{z}, \mathbf{y} \sim N\left(\bar{w}, \frac{\sigma^2}{n}\right). \quad (6.38)$$

The Gibbs sampling algorithm then consists of drawing repeatedly from distributions (6.36), (6.37), and (6.38). Rao–Blackwellization yields good estimates of the posterior means of μ and σ^2 .

6.11.3 Example 6.15: Gibbs Sampling for a Complex Multinomial (*Example 2.4 Continued*)

Consider Example 2.4 of Section 2.4. Let $n = n_O + n_A + n_B + n_{AB}$. Let us denote the observed data by $\check{n} = (n_O, n_A, n_B, n_{AB})$. Suppose you want to do Bayesian estimation of p, q, r with a Dirichlet prior with parameters α, β, γ . The likelihood is, but for a multiplicative constant not involving the parameters,

$$L(p, q, r) = r^{2n_O} (p^2 + 2pr)^{n_A} (q^2 + 2qr)^{n_B} (pq)^{n_{AB}}.$$

The posterior is not particularly elegant, being proportional to

$$r^{2n_O + \gamma - 1} (p^2 + 2pr)^{n_A} (q^2 + 2qr)^{n_B} (p)^{n_{AB} + \alpha - 1} (q)^{n_{AB} + \beta - 1}.$$

It is not easy to deal with this and work out mean, mode, and such useful quantities from the posterior distribution. This is the sort of situation where Gibbs sampling is useful.

Let us introduce two new random variables n_{AA}, n_{BB} , which are unobserved (missing) quantities corresponding to the proportions p^2 and q^2 in the model. Let $n_A = n_{AA} + n_{AO}$, $n_B = n_{BB} + n_{BO}$. Let us write n_{OO} for n_O for the sake of elegant and consistent notation. It is easy to see that if we have observations $\tilde{n} = (n_{OO}, n_{AA}, n_{AO}, n_{BB}, n_{BO}, n_{AB})$, then the likelihood is, but for a multiplicative constant,

$$p^{n_A^+} q^{n_B^+} r^{n_O^+},$$

where

$$\begin{aligned} n_A^+ &= n_{AA} + \frac{1}{2}n_{AB} + \frac{1}{2}n_{AO} \\ n_B^+ &= \frac{1}{2}n_{AB} + n_{BB} + \frac{1}{2}n_{BO}, \\ n_O^+ &= \frac{1}{2}n_{AO} + \frac{1}{2}n_{BO} + n_{OO} \end{aligned}$$

and so the posterior distribution for p, q , and r is easily seen to be Dirichlet with parameters $n_A^+ + \alpha - 1, n_B^+ + \beta - 1, n_O^+ + \gamma - 1$.

This simple solution to this “complete” problem will now be exploited in Gibbs sampling.

Let us observe the following conditional distributions given by the model and the assumed prior:

$$(n_{AA} | \check{n}, p, q, r) \sim \text{Binomial}\left(n_A, \frac{p^2}{p^2 + 2pr}\right), \quad (6.39)$$

$$(n_{BB} | \check{n}, p, q, r) \sim \text{Binomial}\left(n_B, \frac{q^2}{q^2 + 2qr}\right) \quad (6.40)$$

independently.

$$(p, q, r | \check{n}, n_{AA}, n_{BB}) \sim \text{Dirichlet}(n_A^+ + \alpha - 1, n_B^+ + \beta - 1, n_O^+ + \gamma - 1). \quad (6.41)$$

Gibbs sampling for this problem is straightforward—starting from initial estimates for p, q, r , we use random draws from (6.39), (6.40), and (6.41) in turn until “convergence” to

get a random sample from the joint distribution of $(p, q, r, n_{AA}, n_{BB}) \mid \check{n}$. Suppose we have N such independent samples

$$(p^{(j)}, q^{(j)}, r^{(j)}, n_{AA}^{(j)}, n_{BB}^{(j)}).$$

Estimates of posterior mean of p, q, r are obtained by Rao-Blackwellization, namely by

$$\begin{aligned} & \frac{1}{N} \sum_{j=1}^N E(p^{(j)}, q^{(j)}, r^{(j)} \mid \check{n}, n_{AA}^{(i)}, n_{BB}^{(i)}) \\ &= \frac{1}{N} \sum_{j=1}^N (\alpha + n_A^+, \beta + n_B^+, \gamma + n_O^+) / (\alpha + \beta + \gamma + n). \end{aligned}$$

We ran such a Gibbs sampling for this problem with $\alpha = \beta = \gamma = 2$ and the results are presented in Table 6.8 and Figure 6.7.

Table 6.8 Descriptive Statistics from Gibbs Samples of Mean and Rao-Blackwellized (RB) estimates of p, q, r .

Statistic	Mean Estimate			RB Estimate		
	p	q	r	p	q	r
No. of cases	10000	10000	10000	10000	10000	10000
Minimum	0.18845	0.0598	0.5366	0.24943	0.09184	0.61451
Maximum	0.35677	0.1469	0.70758	0.2857	0.10658	0.65646
Median	0.26622	0.09700	0.63578	0.26644	0.09637	0.63719
Mean	0.26727	0.09751	0.63522	0.26608	0.09676	0.63716
SD	0.02206	0.01417	0.02409	0.00662	0.00238	0.00708
Variance	0.00049	0.00020	0.00058	0.00004	0.00001	0.00005
MLE for p, q, r , respectively				0.26444	0.09317	0.64239

6.11.4 Gibbs Sampling Analogs of ECM and ECME Algorithms

Gibbs sampling can be looked upon as an iterative simulation analog of the (partitioned) ECM algorithm (see Section 5.2.2), where the d constraint functions on Ψ are of the form $g_1(\Psi_2, \dots, \Psi_d)$, $g_2(\Psi_1, \Psi_3, \dots, \Psi_d), \dots, g_d(\Psi_1, \dots, \Psi_{d-1})$. This is clearly analogous to the choice of conditioning parameters in Gibbs sampling. In some versions of Gibbs sampling, the variables (Ψ 's) are grouped in order to reduce the number of steps needed at each iteration, when such grouped variables admit a joint conditional distribution simple enough to draw from. This is analogous to integrating out missing data when maximizing a constrained log likelihood function in the ECME algorithm. For instance, suppose we have parameters $\Psi = (\Psi_1^T, \Psi_2^T)^T$. In the ECM algorithm, the distribution of $Z \mid \Psi_1, \Psi_2$ may be used in the E-step and the distributions of $\Psi_1 \mid Z, \Psi_2$ and $\Psi_2 \mid Z, \Psi_1$ may be used in CM-Step 1 and CM-Step 2, respectively. Gibbs sampling analogs use the three distributions for the draws. However, in the ECME algorithm, CM-Step 2 may use the distribution of $\Psi_2 \mid \Psi_1$. Analogous Gibbs sampling uses the draws $Z, \Psi_2 \mid \Psi_1$ and $\Psi_1 \mid Z, \Psi_2$; Liu et al. (1994) give details and results regarding more rapid convergence of

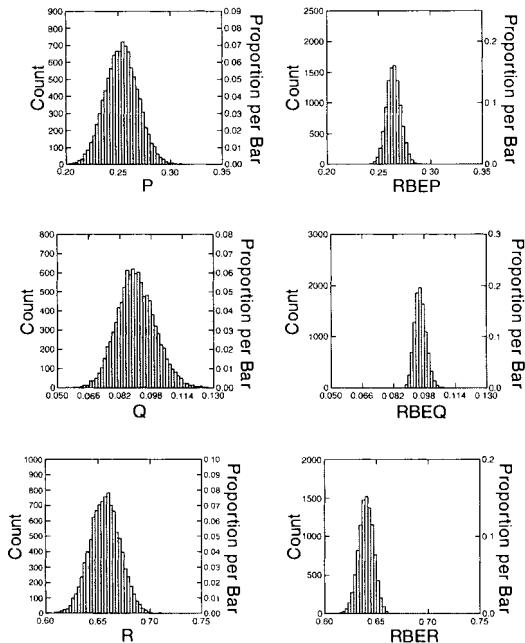


Figure 6.7 Histograms of Mean and Rao-Blackwellized (RB) Estimates of p , q , and r .

collapsed Gibbs sampling (where one or more variables are possible to be integrated out and draws are made from distributions without these) compared to the usual Gibbs sampling. Liu (1994) also shows how to use collapsed Gibbs sampling in the bivariate normal data problem with missing values, for which we discussed the EM algorithm in Example 2.1 of Section 2.2.1.

For the variance components problem discussed in Section 5.9, wherein we presented the EM algorithm and two versions ECME-1 and ECME-2 of the ECME algorithm, we briefly present below Gibbs sampling analogs. They are, in fact, special versions of Gibbs sampling and are actually Data Augmentation algorithms. The parameter vector Ψ for this problem consists of the elements of β , the (distinct) elements of D , and σ^2 ; and the missing-data vector z is $(b_1^T, \dots, b_m^T)^T$. We start with independent improper prior distributions for β , D , and σ^2 with $p(\beta) \propto \text{constant}$, $p(D) \propto |D|^{-1/2}$, and $p(\sigma^2) \propto 1/\sigma$.

6.11.4.1 Gibbs Sampling Analog of EM Algorithm In Gibbs sampling, the missing-data vector z as well as the parameter vector Ψ is sampled. For a Gibbs sampling analog of the EM algorithm, the partition used is z and Ψ . Then the steps of the algorithm are:

Step 1. Given the current draw of Ψ , draw the components b_j of z from the independent conditional distributions $p(b_j | y_j; \beta, D, \sigma^2)$, which are multivariate normal with means and covariance matrices determined from (5.61) and (5.62).

Step 2. Given the draw of z , draw \mathbf{D} , β^T , and σ^2 conditionally independently using a Wishart for \mathbf{D}^{-1} , a p -variate normal for β^T , and a density proportional to χ_{n-1}^2 for σ^2 .

6.11.4.2 Gibbs Sampling Analog of ECME-1 Algorithm In the version of Gibbs sampling analogous to Version 1 of the ECME algorithm, Ψ is partitioned into $(\Psi_1^T, \Psi_2^T)^T$, where $\Psi_1 = \beta$ and Ψ_2 consists of the distinct elements of \mathbf{D} and σ^2 . The algorithm then is

Step 1. Given the current draw of Ψ_2 , draw Ψ_1 from the conditional p -variate normal distribution $p(\beta | \mathbf{y}; \Psi_2)$, and then draw components of z from the conditional distribution given \mathbf{y} .

Step 2. Given Ψ_1 and z , draw Ψ_2 from the conditional distribution using the conditional independence of \mathbf{D} and σ^2 given Ψ_1 as in the previous algorithm.

6.11.4.3 Gibbs Sampling Analog of ECME-2 Algorithm In the version of Gibbs sampling analogous to Version 2 of the ECME algorithm, the partition $(\Psi_1^T, \Psi_2^T)^T$ of Ψ has $\Psi_1 = (\sigma^2, \beta^T)^T$ and Ψ_2 containing the distinct elements of \mathbf{D} . The algorithm then is

Step 1. Given the current draw of Ψ_2 , draw Ψ_1 and draw components of z using the factorization

$$p(\sigma^2, \beta, z | \mathbf{y}; \mathbf{D}) = p(\sigma^2 | \mathbf{y}; \mathbf{D})p(\beta | \mathbf{y}; \sigma^2, \mathbf{D})p(z | \mathbf{y}; \beta, \sigma^2, \mathbf{D}).$$

Step 2. Given Ψ_1 and z , draw \mathbf{D} as in the first algorithm.

It is expected that the approaches of the SAGE and AECM algorithms will give rise to a more flexible formulation of Gibbs sampling incorporating the working parameter concept in the form of a parameter-dependent transformation of the random variable before Gibbs sampling is implemented (see Meng and van Dyk, 1997).

6.12 EMPIRICAL BAYES AND EM

In the following the Empirical Bayes approach (Eggermont and LaRiccia, 2001) to the estimation of a probability density function of a parameter Ψ is formulated as an EM algorithm.

As in a Bayes formulation we consider the parameter Ψ as a random variable with an unknown probability density $p^*(\Psi)$. The density $p^*(\Psi)$ is to be estimated. The complete data set consists of i.i.d. observations

$$(\mathbf{W}_1^T, \Psi_1^T)^T, \dots, (\mathbf{W}_n^T, \Psi_n^T)^T$$

from a probability density $f(\mathbf{w}; \Psi)$, \mathbf{w} being d -dimensional. Missing data are Ψ_1, \dots, Ψ_n . The complete-data log likelihood is

$$\log L_c(p) = \sum_{j=1}^n \log \{f(\mathbf{W}_j; \Psi_j)p(\Psi_j)\} \quad (6.42)$$

which is to be maximized with respect to p over the collection of probability density functions. The solution can be obtained with an EM approach as follows:

Let the initial guess for p^* be p_1 . Consider

$$h(\Psi | \mathbf{w}; p_1) = \frac{f(\mathbf{w}; \Psi)p_1(\Psi)}{\kappa p_1(\mathbf{w})}, \quad (6.43)$$

where

$$\kappa p(\mathbf{w}) = \int_{\Psi} f(\mathbf{w}; \Psi)p(\Psi)d\Psi, \quad \mathbf{w} \in \Re^d. \quad (6.44)$$

Then

$$\begin{aligned} Q(p; p_1) &= E\{\log L_c | \mathbf{W}_1, \dots, \mathbf{W}_n, p_1\} \\ &= \sum_{j=1}^n E[\log\{f(\mathbf{W}_j, \Psi_j)p(\Psi_j)\} | \mathbf{W}_j, p_1], \end{aligned} \quad (6.45)$$

which can be written as

$$\int_{\Psi} p_2(\Psi) \log \frac{p_1(\Psi)}{p(\Psi)} + \text{terms independent of } p, \quad (6.46)$$

where

$$p_2(\Psi) = p_1(\Psi) \frac{1}{n} \sum_{j=1}^n \frac{f(\mathbf{W}_j, \Psi)}{\kappa p_1(\mathbf{W}_j)}. \quad (6.47)$$

The computation of Q can be regarded as an E-step.

In the M-step, the Q function is to be maximized over all probability density functions p . Denoting by KL the Kullback-Leibler divergence between two probability density functions in its arguments and noting that

$$Q(p; p_1) = \text{KL}(p_2, p_1) - \text{KL}(p_2, p) \quad (6.48)$$

the answer to the maximization problem of Q is $p = p_2$. Thus the update for p is of the form

$$p^{(k+1)}(\Psi) = p^{(k)}(\Psi) \frac{1}{n} \sum_{j=1}^n \frac{f(\mathbf{W}_j, \Psi)}{\kappa p^{(k)}(\mathbf{W}_j)}. \quad (6.49)$$

Since

$$Q(p^{(k+1)}; p^{(k)}) - Q(p^{(k)}; p^{(k)}) = \text{KL}(p^{(k+1)}, p^{(k)}), \quad (6.50)$$

$$\log L(p^{(k+1)}) - \log L(p^{(k)}) \geq \text{KL}(p^{(k+1)}, p^{(k)}) \geq 0. \quad (6.51)$$

6.13 MULTIPLE IMPUTATION

In the EM algorithm, the missing values or the latent values are ‘imputed’ in the E-step and complete-data methods are applied on the M-step. Thus the EM algorithm, besides providing MLE’s of parameters, also provides estimates for the missing values. Similarly, in the Data Augmentation algorithm, the imputations $z^{(m)}$ are drawn from appropriate distributions in the I-step. Although these imputed values may be good for the limited purpose of point estimation, using them for other purposes like testing hypotheses may not be suitable.

The method of multiple imputation (MI) is a solution to this problem. Here, for each missing value, a set of M (chosen and fixed) values are simulated from a suitable distribution. The M completed data sets are then analyzed by complete-data methods to obtain M estimates of the Ψ and the covariance matrix of these estimates. These are then used to get a composite estimate of Ψ and its covariance matrix. The multiple imputation method is generally used in a Bayesian framework and the imputed values are drawn from the predictive posterior distribution of the missing values. The imputation can be a step in an iterative scheme like a Data Augmentation algorithm. In fact, the Data Augmentation algorithm can be looked upon as a combination of EM and MI, where the E-step of the EM algorithm is replaced by MI for the missing values and the M-step by MI for Ψ . The technique of MI with applications to survey data analysis is discussed in Rubin (1987) and in Little and Rubin (1989), and the relations between MI and other similar techniques in Rubin (1991).

6.14 MISSING-DATA MECHANISM, IGNORABILITY, AND EM ALGORITHM

In analyzing data with missing values and more generally in incomplete-data problems of the kind we consider in this book, it is important to consider the mechanism underlying the ‘missingness’ or incompleteness. Failure to properly study it and just assuming the missing phenomenon to be purely random and ignorable may vitiate inferences. Rubin (1976) formulates two types of random missingness. If the missingness depends neither on the observed values nor on the missing values, then it is called “Missing Completely at Random” (MCAR). The broader situation where the missingness is allowed to depend on the observed data is called “Missing at Random” (MAR). The situation where missingness does not depend on the observed data is called “Observed at Random” (OAR). MCAR is a special case of MAR, and is a combination of MAR and OAR. The notion of “Parameter Distinctness” (PD) holds if there are no a priori ties between the parameters of the data model and the missingness model. Both MAR and MCAR require that the variable with missing data be unrelated to whether or not a case has missing data on that variable. For example, if those with lower scores are more likely to have missing data on the score variable, the data cannot be MAR or MCAR. When data are not MAR or MCAR, missingness is sometimes said to be “nonignorable”. The difference between MAR and MCAR is based on missingness of data on a variable being related to other variables or not.

Rubin (1976) argues that if the MAR condition holds and the parameter to be estimated and the parameter of the missingness mechanism are distinct, then likelihood inference ignoring the missingness mechanism is valid. Furthermore, under MCAR, sampling inferences without modeling missing-data mechanisms are also valid; see also Little and Rubin (2002).

The phenomenon of incompleteness of data is more general than the missing data phenomenon. Heitjan and Rubin (1991) extend the MAR concept to the general incompleteness situation by defining a notion of “coarsening at random” and show that under this condition, if the coarsening parameters and the parameters to be estimated are distinct, then likelihood and Bayesian inferences ignoring this incompleteness mechanism are valid. Heitjan (1993) gives a number of illustrations of this situation and Heitjan (1994) defines the notion of an “observed degree of coarseness” which helps to extend MCAR to “coarsened completely at random” and illustrates it in a number of applications.

Sarkar (1993) develops sufficiency-based conditions for ignorability of missing data and shows that if these conditions are satisfied, the EM algorithm and Gibbs sampling are

valid. However, in general, the objectives of the analysis need to be taken into account while determining whether missing values can be ignored or not. Estimating parameters only need mild conditions to be satisfied, while imputation of missing values need stronger conditions.

Most of the applications of the EM algorithm and its extensions have been made assuming validity of the inference procedures without checking ignorability conditions. In many situations in practice, the response and incompleteness mechanisms are, in fact, related to the study variables and hence the incompleteness mechanism is nonignorable. In these situations, the inference is biased. A correct way to proceed in these situations is to model the incompleteness mechanisms suitably and formulate a more comprehensive statistical problem. Little (1983a, 1983b, 1993, 1994) and Little and Rubin (2002) discuss quite a few such problems.

In practice, it is rather difficult to determine if data are MAR or MCAR. If only a single variable has missing data, it is often not too difficult to determine if any of the other variables in the data set are associated with this missingness. When data are missing on several variables, determining if other variables are associated with missingness will be complex. Thus determining if even MAR holds for a data set may be quite difficult or impossible just from the data set. Follow-up studies or information from outside the data set will be essential. Generally, MAR may just be a convenient assumption. Although missing data literature suggest many methods for missing data estimation, there are hardly methods available for determining if MAR and MCAR assumptions hold.

When faced with missing data in a multivariate situation, three strategies are generally used:

1. Listwise (casewise) deletion, by which the entire case (data on all the variables in that case) is deleted.
2. Pairwise deletion, by which cases in which observations on one or both of a pair of variables are missing are deleted for purposes of computing a quantity (like covariance or correlation) depending on a pair of variables. Pairwise deletion is frequently used to estimate model parameters in missing data situations. Thus a covariance or a correlation matrix is computed where each element is based on complete data for that pair of variables. This means that different elements of the matrix are computed from different data points. This may result in nonpositive definite matrices, standardized values greater than unity and other problems.
3. Estimation or imputation of missing values: It is clear from the literature that using suitably estimated missing data is better than listwise (casewise) or pairwise deletion, even if data are not MAR.

There are several imputation methods, by which a full data set is created based on the imputation method that fills in data based on information from existing data. Some of the older methods are:

1. Mean imputation, by which the mean from the observed data on a variable is imputed. This has the problem of creating a frequency distribution with a spike in the imputed value, reduction in correlations, underestimation of variance, etc.
2. Regression-based method, by which a (linear) regression model is developed for predicting a missing variable from available variables using complete data on these variables; predicting missing values from such a regression and imputing this for

the missing value. Thus all cases with similar missing value pattern will have the same imputed values causing problems similar to the above, due to unrealistically low levels of noise. A way out is stochastic substitution, by which a random value (generally the regression residual from a randomly selected case) is added to the regression-predicted value. This method is based on the MAR assumption and the assumption that the same model holds for missing and nonmissing cases.

3. Hot-Deck imputation, which imputes new values from “similar” cases. This procedure is believed to produce highly biased coefficients and/or standard errors (Gold and Bentler, 2000).

Relatively newer methods are:

1. Multiple imputation (see Section 6.13).
2. EM Algorithm. The regression method mentioned above is an instance of the EM algorithm in certain cases; see Section 2.2.3. Maximum likelihood estimation (MLE), typically implemented by the EM algorithm, demands fewer statistical assumptions and is generally considered superior to imputation by multiple regression. This assumes missing values are MAR (as opposed to MCAR). It has drawbacks similar to multiple regression imputation such as over-correction.

These methods seem to perform better than the older methods, although the latter are more inconvenient to implement. See Graham and Hofer (2000); Enders and Peugh (2004).

In the case of missing data (MAR and PD) in a sample from the multivariate normal distribution $N_p(\mu, \Sigma)$, the EM algorithm can be used to find MLE's of μ, Σ . They can be used in multivariate methods based on μ, Σ ; for instance, in Discriminant Analysis. However, standard errors computed from such use are biased. One solution to this problem is to specify a nominal sample size m less than actual sample size n ; since it does not account for missingness uncertainty, it is biased. Some suggestions for the choice of m are:

1. The average number of cases: If W_1, \dots, W_p have n_1, \dots, n_p observations, respectively, then

$$m = \frac{1}{p}(n_1 + \dots + n_p).$$

2. Minimum number of cases:

$$m = \min_{1, \dots, p} \{n_1, \dots, n_p\}.$$

3. Minimum of pairwise complete cases.

Heitjan and Basu (1996) discuss the validity of inferences under various types of missingness. They show that

1. If MAR and PD hold, missingness mechanism may be ignored for Bayes/likelihood inferences;
2. If MAR and OAR hold (that is, MCAR holds) missingness mechanism may be ignored for frequentist inference.

They warn that erroneous assumptions of ignorability can result in grossly misleading inferences.

Congdon (2006) discusses missingness and ignorability issues in a Bayesian context and in MCMC computations.

This Page Intentionally Left Blank

SOME GENERALIZATIONS OF THE EM ALGORITHM

7.1 INTRODUCTION

There have been generalizations of the EM algorithm in many different directions. We differentiate these from extensions of the EM algorithm discussed in Chapter 5, although the border between these two terms is fuzzy. In this chapter we discuss some of these generalizations. The method of estimating equations is more general than MLE and leads to the notions of quasi-likelihood and quasi-score as analogs of the likelihood and score functions with similar uses in the context of incomplete data. This leads to the Projection-Solution algorithm. We discuss an estimating equations version of EM in the context of missing or latent data leading to the Expectation-Solution (ES) algorithm. The EM algorithm can be looked upon as a general iterative optimization algorithm, wherein the surrogate function in the form of the expected log likelihood function (the Q -function) is maximized in each iteration. This idea can be extended to other suitable surrogate functions. In the MM (Minorize-Maximize or Majorize-Minimize) algorithm due to Lange (1999, 2004) and Hunter and Lange (2000a, 2000b, 2004) a surrogate function is chosen based on convexity and inequality properties, and optimized; they suggest a large class of such surrogate functions. Neal and Hinton's (1998) Lower Bound Maximization is a similar idea. The EM algorithm is a special case of these classes of algorithms. We also discuss the relation between EM and a few other competing algorithms, like Simulated Annealing.

7.2 ESTIMATING EQUATIONS AND ESTIMATING FUNCTIONS

An Estimating Equation is an equation in observations and parameters which when solved for the parameters in terms of observations gives rise to a parameter estimate. For example the equations,

$$\text{Observed moments} = \text{Corresponding theoretical moments},$$

are a set of estimating equations. Normal equations giving rise to least-squares estimates in linear models are estimating equations. MLEs and M-estimators (robust estimators) are solutions to estimating equations. An equation $h(\mathbf{y}; \Psi) = 0$, where $\mathbf{y} = (y_1, \dots, y_n)^T$ is the vector of observations and Ψ is a vector of parameters, gives rise to an estimating function $h(\mathbf{y}; \Psi)$.

Let y_1, \dots, y_n be i.i.d. with density $f(\cdot; \Psi, \xi)$, and let $h(\mathbf{y}, \Psi)$ be an unbiased estimating function such that $E\{h(\mathbf{y}; \Psi)\} = 0 \quad \forall \Psi, \xi$. Also, let

$$h(\mathbf{y}; \Psi) = \sum_{j=1}^n h(\mathbf{y}_j; \Psi).$$

The optimal unbiased estimating function is defined as the minimizer of

$$S_n = E \left(\frac{h(\mathbf{y}; \Psi)}{E\{\partial h(\mathbf{y}; \Psi)/\partial \Psi\}} \right)^2, \quad (7.1)$$

and when ξ is known, is given by the score function; see Section 3.7 for a discussion on the score statistic or score function. Moreover, if \mathbf{t} is a complete sufficient statistic for ξ for fixed Ψ and the density g of \mathbf{y} factors as

$$g(\mathbf{y}; \Psi, \xi) = f(\mathbf{y} | \mathbf{t}; \Psi) f(\mathbf{t}; \Psi, \xi), \quad (7.2)$$

then the conditional score function

$$\partial f(\mathbf{y} | \mathbf{t}; \Psi) / \partial \Psi$$

is the optimal estimating function for Ψ ; see Liang and Zeger (1995) for a concise treatment of estimating functions. In the right-hand side of (7.2), f is being used a generic symbol for a density function.

7.3 QUASI-SCORE AND THE PROJECTION-SOLUTION ALGORITHM

The method of maximum likelihood estimation (MLE) uses the likelihood function and its derivative, the score function (or score statistic). In the Estimating Function approach, which is more general than MLE, the roles of these functions are played by quasi-likelihood and quasi-score functions, which we define below, following Heyde and Morton (1996).

Let y_j ($j = 1, \dots, n$) be independent responses with regressors or covariates $\mathbf{x}_j = (x_{j1}, \dots, x_{jp})^T$, and put $\mathbf{y} = (y_1, \dots, y_n)^T$. Let the model be

$$\mu_i = E(y_j | \mathbf{x}_j) = h(\mathbf{x}_j^T \Psi). \quad (7.3)$$

For instance $h(y) = y$ is linear regression; $h(y) = [1 + \exp(-y)]^{-1}$ is logistic regression; $h(y) = \exp(y)$ is exponential regression, etc. In classical inference, assumptions are made

on the form of the conditional density $g(y_j \mid \mathbf{x}_j; \Psi)$, likelihood inference is made using the score function, and asymptotic normality is used to make inferences on the regression parameters.

In the Generalized Linear Model (GLM), the systematic part is modeled using $E(y_j) = \mu$ with $h(\mu) = \mathbf{x}_j^T \Psi$, and the random error part is modeled with g belonging to the exponential family,

$$g(y_j; \Psi) \propto \exp\{\gamma_j y_j - b(\gamma_j) + c(y_j)\}$$

for known b, c , and $\gamma_j = \gamma_j(\Psi)$. The score function in the GLM is of the form

$$\sum_{j=1}^n \frac{\partial \mu_j(\Psi)^T}{\partial \Psi} v_j^{-1}(y_j - \mu_j(\Psi)), \quad (7.4)$$

where $v_j = \text{var}(y_j) = \partial^2 b(\gamma)/\partial \gamma^2$. This form used more generally is called the quasi-score function and its integral the quasi-likelihood.

Godambe and Heyde (1987) show that the quasi-score function is optimal with respect to the criterion S_n in expression (7.1) among unbiased estimating equations linear in data.

We now discuss the relationship between quasi-score for complete-data and incomplete-data problems in an EM algorithm set-up with the help of an example. But firstly some notation is needed. In the previous work, we have used $S_c(\mathbf{x}; \Psi)$ to denote the complete-data score statistic and $S(\mathbf{y}; \Psi)$ to denote the incomplete-data score statistic. As seen in Section 3.5, the latter is equal to the conditional expectation of $S_c(\mathbf{x}; \Psi)$ given the incomplete data \mathbf{y} . We now let $S_c^*(\mathbf{x}; \Psi)$ and $S^*(\mathbf{y}; \Psi)$ denote the quasi-score function for the complete and incomplete data, respectively.

In this example, we suppose the complete data \mathbf{x} consist of x_{ij} ($i = 1, 2$; $j = 1, \dots, n$) and the observed data \mathbf{y} consist of $y_j = x_{1j} + x_{2j}$ ($j = 1, \dots, n$). Let

$$\mu_{ij}(\Psi) = E(X_{ij}); \text{ and let } \text{var}(X_{ij}) = \mu_{ij}^2,$$

where X_{ij} is the random variable corresponding to the observation x_{ij} . Consider the estimation of Ψ in the class of linear functions of \mathbf{x} . By (7.4), the quasi-score is

$$S_c^*(\mathbf{x}; \Psi) = \sum_{ij} (x_{ij} - \mu_{ij}) \mu_{ij}^{-2} \frac{\partial \mu_{ij}}{\partial \Psi}.$$

Since we observe only the \mathbf{y} values, we consider only the class of linear functions of \mathbf{y} , a subclass of the linear functions of \mathbf{x} considered for the complete-data problem. The corresponding quasi-score is

$$S^*(\mathbf{y}; \Psi) = \sum_j \{y_j - (\mu_{1j} + \mu_{2j})\} (\mu_{1j}^2 + \mu_{2j}^2)^{-2} \frac{\partial}{\partial \Psi} (\mu_{1j} + \mu_{2j}),$$

the least-squares predictor (projection) of S_c^* , leading to the linear predictor $\hat{\mathbf{x}}$ as

$$\hat{x}_{ij} - \mu_{ij} = \frac{\mu_{ij}^2}{\mu_{1j}^2 + \mu_{2j}^2} [y_j - (\mu_{1j} + \mu_{2j})].$$

Thus

$$S^*(\mathbf{y}; \Psi) = S_c^*(\hat{\mathbf{x}}; \Psi).$$

This projection step is the analog of the EM's E-step. The general operation of projection turns out in this example to be imputation or prediction of the missing data, as happens in

many a case in the E-step in EM algorithms. However, in general, neither the projection nor the E-step is merely imputation of the missing values. The M-step then consists in solving the equation

$$S^*(\mathbf{y}; \Psi) = 0,$$

for the parameter Ψ . This step is called the solution step. The algorithm then consists in iterating these two steps in each cycle. This algorithm is called the Projection-Solution algorithm, an extension of the EM algorithm for the estimating function approach. If the likelihood is available and is used to obtain the score function in the class of estimating functions, then this algorithm reduces to the EM algorithm. In general $E(S) \neq S^*$ unlike in the EM situation; and $E(S)$ may be nonlinear in \mathbf{y} .

In this approach, if $S_c^*(\mathbf{x}; \Psi)$ is the likelihood score for the complete data \mathbf{x} , $S^*(\mathbf{y}; \Psi)$ is not necessarily equal to its conditional expectation (as with the EM algorithm), and this conditional expectation may not be linear in \mathbf{y} . For instance, let $X_{ij} \sim \exp(\mu_{ij})$, the exponential distribution with parameter μ_{ij} . Then

$$E(X_{ij} | y_j; \Psi) = v_j - \frac{y_j \exp(-y_j/v_j)}{1 - \exp(-y_j/v_j)}$$

where

$$v_j^{-1} = \mu_{1j}^{-1} + \mu_{2j}^{-1}.$$

Moreover, $E\{S_c^*(\mathbf{x}; \Psi) | \mathbf{y}\}$ is nonlinear in \mathbf{y} .

In general, suppose for the complete data, $\mathcal{K}_{\mathbf{x}}$ is a family of zero-mean, square integrable functions within which $S_c^*(\mathbf{x}; \Psi)$ is the quasi-score. We adopt $S_c^*(\mathbf{x}; \Psi)$ to obtain $S^*(\mathbf{y}; \Psi)$ in a subclass $\mathcal{K}_{\mathbf{y}}$, usually a suitable linear subspace of $\mathcal{K}_{\mathbf{x}}$, for the incomplete-data problem. In the above example, the space $\mathcal{K}_{\mathbf{y}}$ is the subspace formed by the constraints $y_j = x_{1j} + x_{2j}$, $j = 1, \dots, n$. The quasi-score function $S_c^*(\mathbf{x}; \Psi)$ corresponds in the EM algorithm set-up to the score function $S_c(\mathbf{x}; \Psi)$ for the complete-data log likelihood. The score function based on the incomplete-data log likelihood, which is equal to the conditional expectation of $S_c(\mathbf{x}; \Psi)$ given \mathbf{y} , is approximated iteratively by the gradient of the Q -function,

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log L_c(\Psi) | \mathbf{y} \}$$

and $\Psi^{(k)}$ is the current iterate of Ψ .

In the present approach, the quasi-score function $S^*(\mathbf{y}; \Psi)$ for the incomplete data is obtained as the least-squares predictor of $S_c^*(\mathbf{x}; \Psi)$ for the complete data. More precisely, it is obtained iteratively by solving

$$K(\mathbf{y}; \Psi, \Psi^{(k)}) = \mathbf{0} \tag{7.5}$$

to update $\Psi^{(k)}$ to $\Psi^{(k+1)}$, starting from an initial guess $\Psi^{(0)}$, where $K(\mathbf{y}; \Psi, \Psi^{(k)})$ is defined to be such that

$$E_{\Psi^{(k)}} \| K(\mathbf{y}; \Psi, \Psi^{(k)}) - S_c^*(\mathbf{x}; \Psi) \|^2 = \inf_{G \in \mathcal{K}_{\mathbf{y}}} \{ E_{\Psi^{(k)}} \| G(\mathbf{y}; \Psi) - S_c^*(\mathbf{x}; \Psi) \|^2 \}.$$

If $\Psi^{(k)} \rightarrow \Psi$, then Heyde and Morton (1996) show that in the limit

$$E_{\Psi} \| K(\mathbf{y}; \Psi, \Psi) - S_c^*(\mathbf{x}; \Psi) \|^2 = \inf_{G \in \mathcal{K}_{\mathbf{y}}} \{ E_{\Psi} \| G(\mathbf{y}; \Psi) - S_c^*(\mathbf{x}; \Psi) \|^2 \}.$$

This is the analog of the E-step in EM.

Solving for $\Psi^{(k+1)}$ in (7.5) is the analog of the M-step of the EM algorithm and this is called the solution step of the Projection-Solution algorithm. This is the general form of the Projection-Solution algorithm.

7.4 EXPECTATION-SOLUTION (ES) ALGORITHM

7.4.1 Introduction

The Expectation-Solution (ES) algorithm is a general iterative approach to solving estimating equations involving missing data. The starting point is an estimating equation which may arise from a likelihood, a quasi-likelihood or can be a generalized estimating equation. Thus it is somewhat more general than the quasi-score starting point of the Prediction-Solution algorithm. The E-step is similar to that in the EM algorithm, in the sense it is a conditional expectation computation given observed data and current parameter values; the conditional expectation is that of a statistic of complete data, not necessarily of a sufficient statistic. The analog of the M-step of EM is the solution of this surrogate estimating equation. When the estimating equation arises from the likelihood, this reduces to the EM algorithm. This method leads to a natural way of computing standard errors of estimates. We follow the treatment in Elashoff and Ryan (2004).

Let \mathbf{y} be the observed-data vector, \mathbf{z} the missing data, and $\mathbf{x} = (\mathbf{y}^T, \mathbf{z}^T)^T$ be the complete-data vector. Let Ψ be the d -dimensional parameter vector. Let the complete-data d -dimensional estimating equations be

$$\mathbf{U}_c(\mathbf{x}; \Psi) = \mathbf{0}. \quad (7.6)$$

It is known that solutions to (7.6) are consistent and asymptotically normal if $E(\mathbf{U}_c) = \mathbf{0}$ and the matrix of second derivatives of \mathbf{U}_c with respect to Ψ are positive definite at the true value of Ψ in the complete-data case (Godambe and Kale, 1991).

Suppose that we find a decomposition

$$\begin{aligned} \mathbf{U}_c(\mathbf{x}; \Psi) &= \mathbf{U}_1(\mathbf{y}, \mathbf{S}(\mathbf{x}); \Psi) \\ &= \sum_{j=1}^q \mathbf{a}_j(\Psi) S_j(\mathbf{x}) + \mathbf{b}_{\Psi}(\mathbf{y}) \\ &= \mathbf{A}_{\Psi} \mathbf{S}(\mathbf{x}) + \mathbf{b}_{\Psi}(\mathbf{y}), \end{aligned} \quad (7.7)$$

where \mathbf{a}_j are column d -vectors forming $d \times q$ matrix \mathbf{A} , and \mathbf{b} a column d -vector, \mathbf{S} a q -dimensional function with components S_j . This decomposition need not necessarily be based on the assumption of independent observations. The $\mathbf{S}(\mathbf{x})$ is called a "complete-data summary statistic" (not necessarily a sufficient statistic) and it is q -dimensional. Let

$$\mathbf{h}(\mathbf{y}; \Psi) = E_{\Psi}\{\mathbf{S}(\mathbf{x})|\mathbf{y}\}$$

be a known function and consider the estimating equation

$$\mathbf{U}_0(\mathbf{y}, \Psi) = \mathbf{U}_1(\mathbf{y}, \mathbf{h}(\mathbf{y}; \Psi); \Psi) = \mathbf{0}. \quad (7.8)$$

In view of the linearity of \mathbf{S} , (7.8) is an unbiased estimating equation. The ES iterations are:

E-Step: $\mathbf{S}^{(k)} = \mathbf{h}(\mathbf{y}; \Psi^{(k)})$. This is similar to the E-step in EM; here instead of complete-data sufficient statistics or log likelihood we use complete-data summary statistics that arise from an estimating equation.

S-Step: Solve $\mathbf{U}_1(\mathbf{y}, \mathbf{S}^{(k)}; \Psi) = \mathbf{0}$. This is similar to the M-step in EM; here a surrogate estimating equation is solved.

In the ES algorithm starting from a trial value of $\Psi^{(0)}$ these two steps are iterated in each cycle. If \mathbf{U}_1 arises from the derivative of a log likelihood and \mathbf{S} corresponds to complete-data sufficient statistics, then ES is equivalent to EM.

7.4.2 Computational and Asymptotic Properties of the ES Algorithm

We quote results from Elashoff and Ryan (2004) leading to the method of computing covariance matrices of parameter estimates. The E-step can be rewritten as

$$\mathbf{U}_2(\mathbf{y}; \mathbf{S}; \hat{\boldsymbol{\Psi}}^{(k)}) = \mathbf{S} - \mathbf{h}(\mathbf{y}; \hat{\boldsymbol{\Psi}}^{(k)}) = \mathbf{0}. \quad (7.9)$$

Let

$$\boldsymbol{\delta} = (\boldsymbol{\Psi}, \mathbf{S}), \mathbf{U}(\mathbf{y}; \boldsymbol{\delta}) = (\mathbf{U}_1(\mathbf{y}; \mathbf{S}, \boldsymbol{\Psi}), \mathbf{U}_2(\mathbf{y}; \mathbf{S}, \boldsymbol{\Psi})).$$

Then ES is equivalent to solving $\mathbf{U}(\mathbf{y}; \boldsymbol{\delta}) = \mathbf{0}$. Let $\mathbf{D}(\mathbf{x}, \boldsymbol{\delta})$ be the matrix of partial derivatives of $\mathbf{U}(\mathbf{y}; \boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$. Let \mathbf{V} be the covariance matrix of $\mathbf{U}(\mathbf{y}; \boldsymbol{\delta})$. Then $\sqrt{n}(\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix given by the upper left $d \times d$ submatrix of $\mathbf{G}^{-1}\boldsymbol{\Lambda}(\mathbf{G}^{-1})^T$, where

$$\frac{1}{n}\mathbf{V} \xrightarrow{P} \boldsymbol{\Lambda}, \quad \frac{1}{n}\mathbf{D} \xrightarrow{P} \mathbf{G}, \quad \text{as } n \rightarrow \infty.$$

Let $\mathbf{U}_j(\boldsymbol{\delta})$ be the j th additive independent component of \mathbf{U} . Then $\text{var}(\hat{\boldsymbol{\Psi}})$ can be consistently estimated by the upper left $d \times d$ submatrix of

$$[\mathbf{D}(\mathbf{y}; \boldsymbol{\delta})^{-1} \left(\sum_j \mathbf{U}_j(\boldsymbol{\delta}) \mathbf{U}_j(\boldsymbol{\delta})^T \right) (\mathbf{D}(\mathbf{y}; \boldsymbol{\delta})^{-1})^T]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}}.$$

Alternatively, let

$$h_j(\boldsymbol{\Psi}; \mathbf{y}) = E_{\boldsymbol{\Psi}}\{S_j(\mathbf{x}) \mid \mathbf{y}\}.$$

Let us assume

$$\mathbf{U}_o(\mathbf{y}; \boldsymbol{\Psi}) = \sum_{j=1}^n U_{oj}(\boldsymbol{\Psi}).$$

Denoting by $\dot{\mathbf{D}}(\mathbf{x}; \boldsymbol{\Psi})$ the derivative of $U_o(\mathbf{y}; \boldsymbol{\Psi})$ with respect to $\boldsymbol{\Psi}$, the covariance matrix of $\hat{\boldsymbol{\Psi}}$ is consistently estimated by

$$[\dot{\mathbf{D}}(\mathbf{x}; \boldsymbol{\Psi})^{-1} \left(\sum_{j=1}^n U_{oj}(\boldsymbol{\Psi}) U_{oj}(\boldsymbol{\Psi})^T \right) (\dot{\mathbf{D}}(\mathbf{y}; \boldsymbol{\Psi})^{-1})^T]_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}}$$

and

$$\dot{\mathbf{D}}(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{i=1}^q (\dot{\mathbf{a}}_j h_j + \mathbf{a}_j \dot{h}_j) + \dot{\mathbf{b}}$$

where · means differentiation with respect to $\boldsymbol{\Psi}$.

7.4.3 Example 7.1: Multinomial Example by ES Algorithm (Example 1.1 Continued)

We illustrate the multinomial case of Example 1.1 by the ES algorithm. The complete-data estimating equations are

$$\begin{aligned} \mathbf{U}_c &= \frac{y_{12} + y_4}{\Psi} - \frac{y_2 + y_3}{1 - \Psi} = 0; \\ \mathbf{U}_1 &= \frac{S + y_4}{\Psi} - \frac{y_2 + y_3}{1 - \Psi} = 0. \end{aligned} \quad (7.10)$$

Here $S = y_{12}$ is a missing summary statistic, and

$$U_2 = S - \frac{\Psi}{2 + \Psi}(y_{11} + y_{12}) = 0. \quad (7.11)$$

Iteration of equations (7.10) and (7.11) gives rise to the ES algorithm. This is the same iteration as the EM algorithm and the estimates are the same as in Example 1.1.

To compute the covariance matrix of the estimates, express (7.10) and (7.11) in terms of independent contributions of the 197 observations. Let z_{ij} be the indicator variable which takes the value 1 if observation j is from category i , and is zero otherwise, and put $z_i = \sum_{j=1}^{197} z_{ij}$. Equations (7.10) and (7.11) become

$$U_1 = \sum_{j=1}^n U_{1j} = \sum_{j=1}^n \left(\frac{S}{n\Psi} + \frac{z_{4j}}{\Psi} - \frac{z_{2j} + z_{3j}}{1 - \Psi} \right) = 0.$$

$$U_2 = \sum_{j=1}^n U_{2j} = \sum_{j=1}^n \left[\frac{S}{n} - \frac{\Psi}{2 + \Psi}(z_{1j} + z_{2j}) \right].$$

$$\mathbf{U}_j = \begin{pmatrix} U_{1j} \\ U_{2j} \end{pmatrix} = \begin{pmatrix} \frac{S}{n\Psi} + \frac{z_{4j}}{\Psi} - \frac{z_{2j} + z_{3j}}{1 - \Psi} \\ \frac{S}{n} - \frac{\Psi}{2 + \Psi} z_{1j} \end{pmatrix}.$$

$$\mathbf{D} = \begin{pmatrix} \frac{\partial U_1}{\partial \Psi} & \frac{\partial U_1}{\partial S} \\ \frac{\partial U_2}{\partial \Psi} & \frac{\partial U_2}{\partial S} \end{pmatrix} = \begin{pmatrix} -\frac{S+z_5}{\Psi^2} - \frac{y_2+y_3}{(1-\Psi)^2} & \frac{1}{\Psi} \\ -\frac{2y_1}{(2+\Psi)^2} & 1 \end{pmatrix}.$$

Using (7.8)

$$\mathbf{D}^{-1} \left(\sum_{j=1}^n \mathbf{U}_j \mathbf{U}_j^T \right) (\mathbf{D}^{-1})^T = \begin{pmatrix} 0.002649 & 0.000640 \\ 0.000640 & 0.000213 \end{pmatrix},$$

yielding

$$\text{var}(\hat{\Psi}) = 0.002649.$$

Although the choice of S in this case is clear, in general it is not and criteria are needed to ensure convergence; for more details, see Elashoff and Ryan (2004).

7.5 OTHER GENERALIZATIONS

Simulate-Update Algorithm

Satten and Datta (2000) develop a sequential method for solving estimating equations in missing data problems. Here the score or the estimating function of the estimating equation should be expressible as the expected value of the score function of a suitable complete-data problem with respect to the conditional distribution of the missing data given observed data at current parameter values. In their method, this simulation can be performed by a wide variety of sampling methods including importance sampling methods, rather than this conditional distribution. This step is called the S-step. In the Update Step (U-step), parameters are updated using a closed-form expression, not needing maximization. In this

sequential method the samples generated in a S-step can be used in the subsequent S-step, unlike in say, the MCEM algorithm. The algorithm carries out the S- and U-steps in each cycle until convergence. The sequence of parameter values at the end of each U-step is a stochastic process which converges to the correct solution of the incomplete-data estimating equation almost surely. The U-step is not a maximization step, but a closed-form expression so chosen as to guarantee this convergence.

Expected Estimating Equations

Wang, Huang, Chao, and Jeffcoat (2007) discuss the notion of Expected Estimating Equations in the context of missing data, measurement error, and misclassification, and develop a unified method of estimation. This method is similar to the EM method.

7.6 VARIATIONAL BAYESIAN EM ALGORITHM

We follow the treatment as in Beal and Ghahramani (2003). This algorithm tackles the problem of computing the marginal likelihood of a model, in a Bayesian setup, with latent or missing variables. In this section, we use a generic symbol p to denote probability densities (discrete or continuous), marginal ($p(\cdot)$) or conditional ($p(\cdot | \cdot)$). The marginal likelihood is defined as follows: Let m_i be a model in a set of models with parameters Ψ . By integrating the likelihood $p(\mathbf{y} | \Psi, m_i)$ with respect to a prior distribution $p(\Psi | m_i)$ of Ψ , we obtain the marginal likelihood for the observed data \mathbf{y} under model m_i ,

$$p(\mathbf{y} | m_i) = \int p(\mathbf{y} | \Psi, m_i) p(\Psi | m_i) d\Psi. \quad (7.12)$$

This approach of maximizing the marginal likelihood avoids the overfitting problem associated with maximum likelihood methods. Now, given a prior distribution $p(m_i)$ over the models, the following posterior distributions can be formed,

$$p(m_i | \mathbf{y}) = \frac{p(m_i)p(\mathbf{y} | m_i)}{p(\mathbf{y})}; \quad p(\Psi | \mathbf{y}, m_i) = \frac{p(\mathbf{y} | \Psi, m_i)p(\Psi | m_i)}{p(\mathbf{y} | m_i)}. \quad (7.13)$$

The posterior predictive distribution is then given by the density at a data point \mathbf{y}_o obtained by averaging over the parameters within a model and then over the models as follows,

$$p(\mathbf{y}_o | \mathbf{y}) = \sum_{m_i} \int p(y_o | \Psi, m_i, \mathbf{y}) p(\Psi | m_i, \mathbf{y}) p(m_i | \mathbf{y}) d\Psi. \quad (7.14)$$

As in the set-up of the EM algorithm, let \mathbf{z} denote the latent data or missing values associated with the problem with respect to a formulated complete data problem. In order to maximize the log of the marginal likelihood, a lower bound for it is obtained, which is used as a surrogate maximizer. This lower bound is obtained by considering a distribution $p(\mathbf{z}, \Psi | \mathbf{y}, m)$ with the same support as \mathbf{y} as follows,

$$\begin{aligned} \log p(\mathbf{y} | m_i) &= \log \int p(\mathbf{y}, \mathbf{z}, \Psi | m_i) d\mathbf{z} d\Psi \\ &= \log \int q(\mathbf{z}, \Psi) \frac{p(\mathbf{y}, \mathbf{z}, \Psi | m_i)}{q(\mathbf{z}, \Psi)} d\mathbf{z} d\Psi \\ &\geq \int q(\mathbf{z}, \Psi) \log \frac{p(\mathbf{y}, \mathbf{z}, \Psi | m_i)}{q(\mathbf{z}, \Psi)} d\mathbf{z} d\Psi \end{aligned} \quad (7.15)$$

by Jensen's inequality applied to the concavity of the log function. Maximizing the lower bound with respect to the distribution of $q(\mathbf{z}, \Psi)$ results in $p(\mathbf{z}, \Psi | \mathbf{y}, m_i)$. But the calculation of this requires the normalizing constant, which is a difficult exercise. To solve this problem, a factorized approximation $q(\mathbf{z}, \Psi) = q_{\mathbf{z}}(\mathbf{z})q_{\Psi}(\Psi)$ is used. Then

$$\begin{aligned}\log p(\mathbf{y} | m_i) &\geq q_{\mathbf{z}}(\mathbf{z})q_{\Psi}(\Psi) \log \frac{p(\mathbf{y}, \mathbf{z}, \Psi | m_i)}{q_{\mathbf{z}}(\mathbf{z})q_{\Psi}(\Psi)} d\mathbf{z}d\Psi \\ &= \mathcal{F}_{m_i}(q_{\mathbf{z}}(\mathbf{z}), q_{\Psi}(\Psi), \mathbf{y}),\end{aligned}\quad (7.16)$$

where \mathcal{F} is a functional of the distributions $q_{\mathbf{z}}(\mathbf{z})$ and $q_{\Psi}(\Psi)$. The maximization procedure of \mathcal{F} with respect to its arguments then consists in taking functional derivatives of the lower bound with respect to each of its arguments, holding the other fixed, leading to the iterative algorithm,

$$\cdot q_{\mathbf{z}}^{(k+1)}(\mathbf{z}) \propto \exp\left\{\int \log p(\mathbf{z}, \mathbf{y} | \Psi, m_i) q_{\Psi}^{(k)}(\Psi) d\Psi\right\}; \quad (7.17)$$

$$q_{\Psi}^{(k+1)}(\Psi) \propto p(\Psi | m_i) \exp\left\{\int \log p(\mathbf{z}, \mathbf{y} | \Psi, m_i) q_{\mathbf{z}}^{(k+1)}(\mathbf{z}) d\mathbf{z}\right\}. \quad (7.18)$$

Note the similarity to the E- and M-steps respectively of EM algorithm, the first equation analogous to the 'estimation' of the latent variable and the second updating the parameter value. This has led to this algorithm being called the Variational Bayesian EM (VBEM) algorithm. When the complete-data likelihood belongs to an exponential family with natural parameter vector $a(\Psi)$, Beal and Ghahramani (2003) show that

$$q_{\mathbf{z}}^{(k)}(\mathbf{z}) = p(\mathbf{z} | \mathbf{y}, \mathbf{a}^{(k)}), \quad (7.19)$$

where $\mathbf{a}^{(k)} = E_{q_{\Psi}^{(k)}}\{\mathbf{a}(\Psi)\}$.

On contrasting the VBEM and EM algorithms, the E-step of the EM algorithm proceeds with $\Psi = \Psi^{(k)}$ on the $(k+1)$ th iteration in forming the conditional density of the missing data \mathbf{z} (equivalently, the complete data \mathbf{x}) given the incomplete data \mathbf{y} in order to take the conditional expectation of the complete-data log likelihood given \mathbf{y} . On the M-step of the EM algorithm for MAP estimation, we have from (6.17) that the aim is to maximize $p(\Psi | \mathbf{y}, m_i)$ with respect to Ψ ; that is,

$$\Psi^{(k+1)} = \arg \max_{\Psi} \int q_{\mathbf{z}}^{(k+1)}(\mathbf{z}) \log p(\mathbf{z}, \mathbf{y}, \Psi) d\mathbf{z}.$$

Correspondingly, with VBEM, we have from (7.18) that the updated density for Ψ satisfies

$$q_{\Psi}^{(k+1)}(\Psi) \propto \exp\left[\int q_{\mathbf{z}}^{(k+1)}(\mathbf{z}) \log p(\mathbf{z}, \mathbf{y}, \Psi) d\mathbf{z}\right].$$

Thus the EM for MAP is a special case of VBEM where on the $(k+1)$ th iteration $q_{\Psi}^{(k)}(\Psi)$ is chosen to be the Dirac delta function $\delta(\Psi - \Psi^{(k)})$, which puts mass one at the point $\Psi = \Psi^{(k)}$; see Beal and Ghahramani (2003).

Recently, Wang and Titterington (2006) have proposed a generalized iterative algorithm for calculating variational Bayes estimates for a normal mixture model. They showed theoretically that the estimator produced converges locally to the MLE estimator at the rate of $O(1/n)$ in the sample limit.

7.7 MM ALGORITHM

7.7.1 Introduction

The MM algorithm stands for Majorization–Minimization or Minorization–Maximization algorithm. The majorization or the minorization step is a generalization of the E-step in the EM algorithm and the minimization or maximization is the analog of the Maximization step of the EM algorithm.

We now describe the majorization-minorization version of the MM approach in search of the minimum of a function f . We follow the treatments in Lange, Hunter, and Yang (2000a, 2000b), Hunter and Lange (2000a, 20004), Hunter (2003), and Lange (2004).

Let $\Psi^{(k)}$ represent a fixed value of the parameter Ψ and $h(\Psi; \Psi^{(k)})$ a real-valued function of Ψ . Then $h(\Psi; \Psi^{(k)})$ is said to majorize a real-valued function $f(\Psi)$ at $\Psi^{(k)}$ if

$$h(\Psi; \Psi^{(k)}) \geq f(\Psi) \quad \forall \Psi, \quad (7.20)$$

$$h(\Psi^{(k)}; \Psi^{(k)}) = f(\Psi^{(k)}). \quad (7.21)$$

$h(\Psi; \Psi^{(k)})$ is said to minorize $f(\Psi)$ if $(-h(\Psi; \Psi^{(k)}))$ majorizes $(-f(\Psi))$ at $\Psi^{(k)}$.

Here $\Psi^{(k)}$ represents the current iterate in search of $f(\Psi)$. Instead of optimizing the actual function $f(\Psi)$, we optimize a surrogate function $h(\Psi; \Psi^{(k)})$. Thus

$$h(\Psi^{(k+1)}; \Psi^{(k)}) \leq h(\Psi^{(k)}; \Psi^{(k)}). \quad (7.22)$$

This is the gist of the MM algorithm.

The reason why the MM algorithm works is as follows:

$$\begin{aligned} f(\Psi^{(k+1)}) &= h(\Psi^{(k+1)}; \Psi^{(k)}) + f(\Psi^{(k+1)}) - h(\Psi^{(k+1)}; \Psi^{(k)}) \\ &\leq h(\Psi^{(k)}; \Psi^{(k)}) + f(\Psi^{(k)}) - h(\Psi^{(k)}; \Psi^{(k)}) \\ &= f(\Psi^{(k)}), \end{aligned}$$

because by (7.22)

$$h(\Psi^{(k+1)}; \Psi^{(k)}) \leq h(\Psi^{(k)}; \Psi^{(k)}),$$

and by (7.20)and (7.21)

$$f(\Psi^{(k+1)}) - h(\Psi^{(k+1)}; \Psi^{(k)}) \leq f(\Psi^{(k)}) - h(\Psi^{(k)}; \Psi^{(k)}) = 0.$$

Thus, in a minimization problem, with a majorizer h , MM ensures $f(\Psi^{(k+1)}) \leq f(\Psi^{(k)})$ and similarly in a maximization problem, with a minorizer h , MM ensures that $f(\Psi^{(k+1)}) \geq f(\Psi^{(k)})$. Thus MM stands for Majorization–Minimization or Minorization–Maximization. Heiser (1995) shows that the E-step of the EM algorithm is a minimization step.

In an interesting article Hunter (2003) discusses the geometry of the EM algorithm, which elucidates the ideas above.

Application of the MM algorithm to quantile regression without sorting can be found in Hunter and Lange (2000a, 2004).

Figure 7.1 gives an example of a rather complicated likelihood function and a smoother minorizer, as defined above. The minorizer lies below the likelihood and is tangential to it at the MLE.

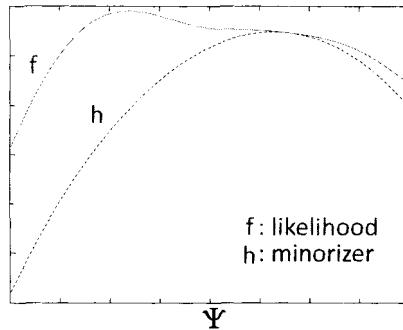


Figure 7.1 A likelihood function f and its minorizer h .

The majorization relation is closed under sums, nonnegative products, limits, composition with increasing functions. In EM we maximize a surrogate function, the expected value of log likelihood given data and current parameter values, whereas in MM, there is a wider choice of the surrogate function. For MM, it is enough to decrease (increase) the surrogate function over iterations, and need not minimize (maximize). The EM analog of such a procedure is the GEM. MM, like EM, substitutes a simple optimization problem for a difficult one, by means of a surrogate function. The advantages are as pointed out by Lange (2004). It

1. avoids large matrix inversions;
2. linearizes the problem;
3. separates parameters;
4. deals with constraints gracefully;
5. turns a nondifferentiable problem into a smooth problem.

The price paid for these advantages is iteration.

7.7.2 Methods for Constructing Majorizing/Minorizing Functions

Majorizing–Minorizing functions can be obtained through suitable inequalities. The above references suggest a host of inequalities in such contexts. Some of them are:

1. Arithmetic Mean \geq Geometric Mean;

2. Cauchy-Schwarz Inequality;

3. Jensen's Inequality;

4. Minimization via Supporting Hyperplane;

Any linear function tangent to a graph of a convex function is a minimizer at point of tangency. Let $\kappa(\cdot)$ be a convex and differentiable function. Then

$$\kappa(\Psi) \geq \kappa(\Psi^{(k)}) + \nabla \kappa(\Psi^{(k)})^T (\Psi - \Psi^{(k)}),$$

with ∇ being the gradient operator.

5. Majorization via Convexity;

Convexity:

$$\kappa(\sum_i \alpha_i t_i) \leq \sum_i \alpha_i \kappa(t_i); \quad \alpha_i \geq 0, \sum_i \alpha_i = 1.$$

Let \mathbf{w} , Ψ , and $\Psi^{(k)}$ be vectors. Then if

$$\alpha_i t_i = x_i (\Psi_i - \Psi_i^{(k)}) + \mathbf{w}^T \Psi^{(k)},$$

$$\kappa(\mathbf{w}^T \Psi) \leq \sum_i \kappa(x_i [\Psi_i - \Psi_i^{(k)}] + \mathbf{w}^T \Psi^{(k)}).$$

If now

$$t_i = \frac{\mathbf{w}^T \Psi^{(k)} \Psi_i}{\Psi_i^{(k)}}, \quad \alpha_i = \frac{x_i \Psi_i^{(k)}}{\mathbf{w}^T \Psi^{(k)}},$$

then

$$\kappa(\mathbf{w}^T \Psi) \leq \sum_i \frac{x_i \Psi_i^{(k)}}{\mathbf{w}^T \Psi^{(k)}} \kappa\left(\frac{\mathbf{w}^T \Psi^{(k)} \Psi_i}{\Psi_i^{(k)}}\right).$$

6. Majorization via Quadratic Upper Bound;

Let $\kappa(\Psi)$ be convex, twice differentiable, have bounded curvature. Then majorization by quadratic function with sufficiently high curvature and tangent to $\kappa(\Psi)$ at $\Psi^{(k)}$ is done as follows:

If M is positive definite and $M - \nabla^2 \kappa(\Psi)$ is nonnegative definite $\forall \Psi$, then

$$\kappa(\Psi) \leq \kappa(\Psi^{(k)}) + \nabla \kappa(\Psi^{(k)})^T (\Psi - \Psi^{(k)}) + \frac{1}{2} (\Psi - \Psi^{(k)})^T M (\Psi - \Psi^{(k)}).$$

For scalar Ψ ,

$$\frac{1}{\Psi} \leq \frac{1}{\Psi^{(k)}} - \frac{\Psi - \Psi^{(k)}}{\Psi^{(k)2}} + \frac{(\Psi - \Psi^{(k)})^2}{c^3}$$

for

$$0 < c \leq \min(\Psi, \Psi^{(k)}).$$

7.7.3 Example 7.2: MM Algorithm for the Complex Multinomial (Example 1.1 Continued)

In Example 1.1, we want to maximize the following log likelihood function

$$\log L(\Psi) = 2n_0 \log r + n_A \log(p^2 + 2pr) + n_B \log(q^2 + 2qr) + 2n_{AB} \log(pq)$$

subject to the equality constraint $p + q + r = 1$ and nonnegativity constraints $p, q, r \geq 0$. The terms $n_A \log(p^2 + 2pr)$, $n_B \log(q^2 + 2qr)$ make the maximization difficult and we tackle it by using the convexity of the function $(-\log x)$ and the minorization

$$\begin{aligned} \log(p^{(k)2} + 2p^{(k)}r^{(k)}) &\geq \frac{p^{(k)2}}{p^{(k)2} + 2p^{(k)}r^{(k)}} \log\left(\frac{p^{(k)2} + 2p^{(k)}r^{(k)}}{p^{(k)2}} p^{(k)2}\right) \\ &\quad + \frac{2p^{(k)}r^{(k)}}{p^{(k)2} + 2p^{(k)}r^{(k)}} \log\left(\frac{p^{(k)2} + 2p^{(k)}r^{(k)}}{p^{(k)2}} 2p^{(k)}r^{(k)}\right). \end{aligned}$$

There is a similar minorization of $\log(q^2 + 2qr)$.

Using the notation,

$$n_{AA}^{(k)} = n_A \frac{p^{(k)2}}{p^{(k)2} + 2p^{(k)}r^{(k)}},$$

$$n_{AO}^{(k)} = n_A \frac{2p^{(k)}r^{(k)}}{p^{(k)2} + 2p^{(k)}r^{(k)}},$$

and similar notations $n_{BB}^{(K)}$ and $n_{BO}^{(k)}$, the surrogate function that could be maximized is

$$\begin{aligned} h(p, q, r; p^{(k)}, q^{(k)}, r^{(k)}) &= n_{AA}^{(k)} \log p^2 + n_{AO}^{(k)} \log(2pr) \\ &\quad + n_{BB}^{(k)} \log q^2 + n_{BO}^{(k)} \log(2qr) \\ &\quad + n_{AB} \log(2pq) + n_O \log r^2. \end{aligned}$$

This is the minorization step.

The maximization step is accomplished by using a Lagrange multiplier and finding a stationary point of

$$h^*(p, q, r) = h(p, q, r; p^{(k)}, q^{(k)}, r^{(k)}) + \lambda(p + q + r - 1).$$

This involves equating to zero the following derivatives,

$$\begin{aligned} \partial h^*(p, q, r)/\partial p &= \frac{2n_{AA}^{(k)} + n_{AO}^{(k)} + n_{AB}}{p} + \lambda, \\ \partial h^*(p, q, r)/\partial q &= \frac{2n_{BB}^{(k)} + n_{BO}^{(k)} + n_{AB}}{q} + \lambda, \\ \partial h^*(p, q, r)/\partial r &= \frac{2n_{AO}^{(k)} + n_{BO}^{(k)} + 2n_O}{r} + \lambda, \\ \partial h^*(p, q, r)/\partial \lambda &= p + q + r - 1. \end{aligned}$$

This provides the updates

$$p^{(k+1)} = \frac{n_A^+}{n}; \quad q^{(k+1)} = \frac{n_B^+}{n}; \quad r^{(k+1)} = \frac{n_O^+}{n},$$

exactly as with the EM.

Hunter (2004) gives an application of the MM algorithm for maximum likelihood estimation of parameters in a generalized version of the Bradley-Terry model used in paired comparison studies.

7.8 LOWER BOUND MAXIMIZATION

Neal and Hinton (1998) give an explanation of the EM algorithm in terms of optimization of the surrogate function as in the MM algorithm above, which provides a nice insight into the EM algorithm. We follow the treatment in Dellaert (2002), and Minka (1998) and explain the idea with the finite mixture example. We have data \mathbf{y} from the mixture

$$g(\mathbf{y}; \boldsymbol{\Psi}) = \sum_{i=1}^g \pi_i f_i(\mathbf{y}; \boldsymbol{\Psi}).$$

We want to find

$$\Psi^* = \arg \max_{\Psi} \log f(\mathbf{y}; \Psi) = \arg \max_{\Psi} \log \sum_{i=1}^g \pi_i f_i(\mathbf{y}; \Psi).$$

Now

$$\log g(\mathbf{y}; \Psi) = \log \sum_{i=1}^g \pi_i f_i(\mathbf{y}; \Psi) = \log \sum_{i=1}^g p(i; \Psi^{(k)}) \frac{\pi_i f_i(\mathbf{y}; \Psi)}{p(i; \Psi^{(k)})},$$

where $p(i; \Psi^{(k)})$ is an arbitrary probability distribution over $\{1, \dots, g\}$, depending on the current value $\Psi^{(k)}$ of Ψ (or the initial guess for $k = 0$). Now we compute a lower bound $B(\Psi; \Psi^{(k)})$ for $\log g(\mathbf{y}; \Psi)$, using $\Psi^{(k)}$ as follows. By Jensen's inequality,

$$\log \sum_{i=1}^g p(i | \Psi^{(k)}) \frac{\pi_i f_i(\mathbf{y}; \Psi)}{p(i; \Psi^{(k)})} \geq \sum_{i=1}^g p(i; \Psi^{(k)}) \log \frac{\pi_i f_i(\mathbf{y}; \Psi)}{p(i; \Psi^{(k)})} \quad (7.23)$$

$$= B(\Psi; \Psi^{(k)}), \quad (7.24)$$

say. Now we find an optimal lower bound with respect to the distribution $p(i; \Psi^{(k)})$ by maximizing $B(\Psi^{(k)}; \Psi^{(k)})$. The restriction of $p(i; \Psi)$ being a probability distribution means $\sum_{i=1}^g p(i; \Psi^{(k)}) = 1$. The objective function to be maximized with a Lagrange multiplier λ is

$$G(p) = \lambda \left\{ \sum_{i=1}^g p(i; \Psi^{(k)}) - 1 \right\} + \sum_{i=1}^g [p(i; \Psi^{(k)}) \{ \log \pi_i + \log f_i(\mathbf{y}; \Psi^{(k)}) - \log p(i; \Psi^{(k)}) \}].$$

Differentiating G with respect to p and equating it to 0, and using the constraint, we obtain

$$\begin{aligned} p(i; \Psi^{(k)}) &= \pi_i f_i(\mathbf{y}; \Psi^{(k)}) / g(\mathbf{y}; \Psi^{(k)}) \\ &= \tau_i(\mathbf{y}; \Psi^{(k)}), \end{aligned} \quad (7.25)$$

where

$$\tau_i(\mathbf{y}; \Psi) = \text{pr}_{\Psi} \{ i | \mathbf{y} \}.$$

From (7.24) and (7.25),

$$\begin{aligned} B(\Psi^{(k)}; \Psi^{(k)}) &= \sum_{i=1}^g \tau_i(\mathbf{y}; \Psi^{(k)}) \log \frac{\pi_i f_i(\mathbf{y}; \Psi^{(k)})}{\tau_i(\mathbf{y}; \Psi^{(k)})} \\ &= \log g(\mathbf{y}; \Psi^{(k)}), \end{aligned}$$

and

$$B(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log g(\mathbf{y}; \Psi) | \mathbf{y} \} - \sum_{i=1}^g \tau_i(\mathbf{y}; \Psi^{(k)}) \log \tau_i(\mathbf{y}; \Psi^{(k)}), \quad (7.26)$$

where the first term on the right-hand side of (7.26) is the Q -function computed on the E-step of the EM algorithm,

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log g(\mathbf{y}; \Psi) | \mathbf{y} \};$$

that is, it is the conditional expectation of the complete-data log likelihood using the current value for Ψ . The other term on the right-hand side of (7.26) does not depend on Ψ , and so maximizing the optimal lower bound is the same as the usual M-step of the EM algorithm. If we consider the log posterior instead of the log likelihood in order to get a MAP estimate, then there will be an additional term in the function to be maximized, leading to the **M-Step**:

$$\Psi^{(k+1)} = \arg \max_{\Psi} \{Q(\Psi; \Psi^{(k)}) + \log p(\Psi)\},$$

where $p(\Psi)$ is the prior.

7.9 INTERVAL EM ALGORITHM

7.9.1 The Algorithm

We follow the treatment in Wright and Kennedy (2000).

The object of the Interval EM algorithm is to locate all the stationary points of the log likelihood in a given region of the parameter space. The algorithm is a combination of the bisection algorithm and the EM idea of obtaining an ‘enclosure’ of the gradient of the log likelihood. We use interval functions which are interval-valued functions of interval arguments. An interval function $R(\mathbf{w}_1^I, \dots, \mathbf{w}_n^I)$ of intervals $\mathbf{w}_1^I, \dots, \mathbf{w}_n^I$ is said to be an extension or enclosure of a function $r(w_1, \dots, w_n)$ if $R([w_1, w_1], \dots, [w_n, w_n]) = r(w_1, \dots, w_n)$ for all w_1, \dots, w_n .

We know that the Q -function of the EM algorithm has a gradient which is equal to the gradient of the log likelihood at stationary points of the log likelihood. The interval EM algorithm constructs interval vectors which enclose values of the gradient of the Q -function. The method gradually eliminates regions of the parameter space not containing stationary points, by finding a region which encloses the range of the gradient of the log likelihood in a given region. An initial box (maybe a large one) is specified in the parameter space containing a stationary point and it is gradually made smaller by bisection, evaluating the enclosure of the log likelihood over each box; boxes that do not contain a stationary point are discarded. The algorithm stops when the box is sufficiently small. The list of all such boxes contain all the stationary points in the initial region.

Let Ω be the parameter space, Ψ the parameter vector. The Interval EM method generates a sequence of intervals $\Psi_1^I, \dots, \Psi_k^I$ in Ω with enclosures $Q(\Psi; \Psi_0^I), Q(\Psi; \Psi_1^I), \dots, Q(\Psi; \Psi_k^I)$ of functions $Q(\Psi; \Psi_0), Q(\Psi; \Psi_1), \dots, Q(\Psi; \Psi_k)$ so that $Q(\Psi; \Psi_0) \in Q(\Psi; \Psi_0^I)$ where $\Psi_j \in \Psi_j^I$ for each j . The interval Ψ_{j+1}^I contains at least one value of Ψ_{j+1} maximizing a $Q(\Psi; \Psi_j)$ for at least one $\Psi_j \in \Psi_j^I \subset \Omega$. Interval EM step consists in moving from Ψ_j^I to Ψ_{j+1}^I .

The algorithm can be implemented by a full bisection search of each of a list of boxes. A sequence of bisection of a box is carried out at each of its coordinates. A box is discarded if it does not contain a 0 in at least one direction of the enclosure of the gradient. This is repeated until the diameter of every box is small.

7.9.2 Example 7.3: Interval-EM Algorithm for the Complex Multinomial (Example 2.4 Continued)

Let us consider Example 2.4 of Multinomial with complex cell structure. Here the Q -function $Q(p, q; p^{(k)}, q^{(k)})$ is given by

$$\begin{aligned}
Q(p, q; p^{(k)}, q^{(k)}) &= \frac{182}{1 + 2(1 - p^{(k)} - q^{(k)})/p^{(k)} + 199} \log p \\
&\quad + \frac{60}{1 + 2(1 - p^{(k)} - q^{(k)})/q^{(k)} + 77} \log q \\
&\quad + \left(594 - \frac{182}{1 + 2(1 - p^{(k)} - q^{(k)})/p^{(k)}}\right) \log r \\
&\quad - \frac{60}{1 + 2(1 - p^{(k)} - q^{(k)})/q^{(k)}} \log r,
\end{aligned}$$

where $r = 1 - p - q$.

The only stationary point located inside

$$(p_0, q_0) = ([0.00001, 0.45], [0.00001, 0.45])$$

is located inside

$$([0.2644443138466694, 0.2644443138466706], [0.09316881181568122, 0.09316881181568200]).$$

7.10 COMPETING METHODS AND SOME COMPARISONS WITH THE EM ALGORITHM

7.10.1 Introduction

It is generally acknowledged that the EM algorithm and its extensions are the most suitable for handling the kind of incomplete-data problems discussed in the book. However, quite a few other methods are advocated in the literature as possible alternatives to EM-type algorithms. Some of these like Simulated Annealing are general-purpose optimization methods, some like the Delta algorithm are special-purpose ML methods, and others like the Image Space Reconstruction Algorithm (ISRA) are algorithms meant for special applications. Some comparisons of the performance of these algorithms with the EM algorithm have been reported for specific problems like the normal mixture resolution problem. We present a brief review of these methods and the comparisons reported. Some other methods and comparisons not reported here can be found in Davenport, Pierce, and Hathaway (1988), Campillo and Le Gland (1989), Celeux and Diebolt (1990, 1992), Fairclough, Pierni, Ridgway, and Schwertman (1992), Yuille, Stolorz, and Utans (1994), and Xu and Jordan (1996).

7.10.2 Simulated Annealing

Simulated Annealing (SA) was introduced by Kirkpatrick, Gelatt, and Vecchi (1983) and Cerný (1985), but in its original form, it dates back to Pincus (1968, 1970); see Ripley (1990) on this point. This is a stochastic-type technique for global optimization using the Metropolis algorithm based on the Boltzmann distribution in statistical mechanics. It is an algorithm in the MCMC family.

To minimize a real-valued function $h(\Psi)$ on a compact subset Ω of \mathbb{R}^d , SA generates an inhomogeneous Markov process on Ω depending on the positive time-parameter T (called

the *temperature* in accordance with the physical analogy and also to differentiate it from the statistical parameter Ψ). This Markov process has the following Gibbs distribution as its unique stationary distribution,

$$p_T(\Psi) = \frac{\exp(-h(\Psi)/T)}{\int_{\Omega} \exp(-h(\Psi)/T) d\Psi}, \quad \Psi \in \Omega. \quad (7.27)$$

In the limit as T tends to 0 from above, the stationary distribution tends to a distribution concentrated on the points of global minima of h . In a homogeneous version of the algorithm, a sequence of homogeneous Markov chains is generated at decreasing values of T .

The general algorithm is as follows:

- Step 1.** Choose an initial value for T , say T_0 and an initial value Ψ_0 of Ψ with $h(\Psi_0) = h_0$.
- Step 2.** Select a proposed point Ψ_1 at random from a neighborhood of Ψ_0 and calculate the corresponding h -value.
- Step 3.** Compute $\Delta_1 = h(\Psi_1) - h(\Psi_0)$. If $\Delta_1 \leq 0$, move to the new point Ψ_1 . Otherwise draw a value u from the uniform distribution over $[0, 1]$. Then accept the new point Ψ_1 if $u \leq \exp(-\Delta_1/T)$, i.e., $\exp(-\Delta_1/T)$ is the probability of acceptance.
- Step 4.** Repeat Steps 2 and 3 after updating the appropriate quantities, until an equilibrium has been reached by the application of a suitable stopping rule.
- Step 5.** Lower the temperature according to an “annealing schedule” and start at Step 2 with the equilibrium value at the previous T as the starting value. Again a suitable stopping rule between temperatures is used to decide to stop the algorithm giving the solution to the minimization problem.

The stopping rules, the annealing schedule, and the methods of choosing the initial and next values constitute the “cooling schedule”. It is well known that the behavior of the SA approach to optimization is crucially dependent on the choice of cooling schedule. The literature gives some standard methods of arriving at these. The reader is referred to Ripley (1988) and Weiss (1989) for an introduction to SA, and to van Laarhoven and Aarts (1987); Aarts and van Laarhoven (1989); and Bertsimas and Tsitsiklis (1993) for general reviews of the literature. Geman and Geman (1984) adopted the SA approach for MAP estimation in image analysis, using the Gibbs sampler to simulate (7.27).

7.10.3 Comparison of SA and EM Algorithm for Normal Mixtures

Ingrassia (1991, 1992) compares SA and the EM algorithm for univariate normal mixtures by simulation. The study of Ingrassia (1992) is an improved version of Ingrassia (1991) as it chooses a constrained formulation of the mixture problem, uses an improved cooling schedule for the SA algorithm, and considers several normal mixtures. The criteria for comparison are distances like the Kullback–Leibler divergence measures between distribution functions—between the one used for simulation (the true one) and the one obtained by the algorithm. It is found that though neither algorithm overwhelmingly outperforms the other, SA performs more satisfactorily in many cases. However, SA is much slower than the EM algorithm. Though SA gives estimates closer to true values compared to the EM algorithm, in a number of these cases, the value of the likelihood is greater at the EM solution. Thus the superiority of SA is not so definitive.

Brooks and Morgan (1995) make a comparison between traditional methods such as quasi-Newton and a sequential Quadratic Programming algorithm with SA as well as with hybrids of SA and these traditional methods (where the SA solution is used as a starting point for the traditional method). They find that the traditional methods repeatedly get stuck at points associated with singularities. They conclude that the hybrid algorithms perform well and better than either the traditional methods or SA; for comparisons of the EM algorithm and other methods for the normal mixture problems, see also Everitt (1984), Davenport et al. (1988), Lindsay and Basak (1993), and Aitkin and Aitkin (1996).

7.11 THE DELTA ALGORITHM

Jørgensen (1984) introduces the Delta algorithm. It is a generalization of the scoring method, is a modification of the Newton-Raphson method, and has an interpretation as an iterative weighted least-squares method.

Let the log likelihood be of the form $\log L(\boldsymbol{\mu}(\boldsymbol{\Psi}))$, where $\boldsymbol{\Psi}$ is a $d \times 1$ vector of parameters to be estimated and $\boldsymbol{\mu}$ is a $n \times 1$ vector function of $\boldsymbol{\Psi}$, n being the sample size. For instance, the mean $\boldsymbol{\mu}$ could be of the form $\mathbf{X}\boldsymbol{\Psi}$, where \mathbf{X} is a $n \times d$ (design) matrix, as in the case of a linear model. In the Delta algorithm, the matrix of second derivatives usually used in the Newton-Raphson algorithm is replaced by an approximation of the form

$$-\partial^2 \log L(\boldsymbol{\mu})/\partial\boldsymbol{\Psi}\partial\boldsymbol{\Psi}^T \approx \mathbf{X}^T(\boldsymbol{\Psi})\mathbf{K}(\boldsymbol{\mu})\mathbf{X}(\boldsymbol{\Psi}), \quad (7.28)$$

where $\mathbf{K}(\boldsymbol{\mu})$ is a symmetric positive definite matrix, called the weight matrix. Here $\mathbf{X}(\boldsymbol{\Psi}) = \partial\boldsymbol{\mu}/\partial\boldsymbol{\Psi}^T$ is called the local design matrix and $\mathbf{K}(\boldsymbol{\mu})$ is to be chosen suitably. The name Delta method is due to the similarity of equation (7.28) with the way asymptotic covariance matrices are obtained when parameters are transformed, using the traditional delta method.

If $\mathbf{K}(\boldsymbol{\mu})$ is the expected information matrix for $\boldsymbol{\mu}$, then the Delta method is the scoring method.

Böhning and Lindsay (1988) and Böhning (1993) discuss a method called the Lower Bound algorithm, which is a special case of the Delta algorithm. Here if the Hessian is bounded below in the sense of Lowener ordering (defined by $\mathbf{A} \geq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is nonnegative definite), that is,

$$\partial^2 \log L(\boldsymbol{\mu})/\partial\boldsymbol{\Psi}\partial\boldsymbol{\Psi}^T \geq \mathbf{B}$$

for all $\boldsymbol{\Psi} \in \Omega$, where \mathbf{B} is a symmetric, negative definite matrix not depending on $\boldsymbol{\Psi}$, then $(-\mathbf{B})$ is used as the \mathbf{K} matrix. Such a bound can be used to create a monotonically convergent modification of the Newton-Raphson algorithm; see our discussion in Section 1.3.2. Böhning and Lindsay (1988) give examples of such likelihoods and call them Type I likelihoods.

We assume that the matrix $\mathbf{X}(\boldsymbol{\Psi})$ has full rank for all $\boldsymbol{\Psi}$. Differentiability conditions to get the score vector and information matrices are assumed. Note that

$$\partial \log L(\boldsymbol{\mu})/\partial\boldsymbol{\Psi} = \mathbf{X}^T(\boldsymbol{\Psi})S(\boldsymbol{\mu}),$$

where $S(\boldsymbol{\mu}) = \partial \log L(\boldsymbol{\mu})/\partial\boldsymbol{\mu}$. Let $\mathbf{X}(\boldsymbol{\Psi}^{(k)})$ be denoted by $\mathbf{X}^{(k)}$. Let $\boldsymbol{\Psi}^{(0)}$ be the starting value of the parameter for the Delta algorithm. The algorithm is defined by

$$\boldsymbol{\Psi}^{(k+1)} = \boldsymbol{\Psi}^{(k)} + a^{(k)}\boldsymbol{\Psi}^{(k)},$$

where

$$\Psi^{(k)} = (\mathbf{X}^{(k)T} \mathbf{K} \mathbf{X}^{(k)})^{-1} \mathbf{X}^{(k)T} S^{(k)}$$

and $a^{(k)}$ is a number called the step-length chosen to make

$$\log L(\boldsymbol{\mu}^{(k+1)}) \geq \log L(\boldsymbol{\mu}^{(k)}).$$

The step-length $a^{(k)}$ can be taken to be 1 (then called the unit-step Delta algorithm), but it is recommended that a search be made to choose $a^{(k)}$ so as to increase the value of $\log L(\boldsymbol{\mu})$.

Jørgensen (1984) discusses various aspects of the algorithm, including choice of starting values, choice of the weight matrix, computation of standard errors of the estimates and implementation in GLIM. He also gives a few illustrative examples.

Jørgensen (1984) suggests that for computing the MLE for an incomplete-data problem via the EM algorithm, on the M-step one iteration of the Delta algorithm be used with $S^{(k)}$ replaced by its current conditional expectation given the observed data. He shows that a step-length chosen to increase the Q -function also increases the incomplete-data log likelihood. He also shows that the weight matrix may be chosen based on Q or as $\mathcal{I}(\boldsymbol{\mu}^{(k)})$ or $\mathcal{I}(\boldsymbol{\mu}^{(k)}; \mathbf{y})$.

7.12 IMAGE SPACE RECONSTRUCTION ALGORITHM

In Section 2.5, we considered the EM algorithm for the PET problem. Daube-Witherspoon and Muehllehner (1986) modify the EM algorithm for this problem by replacing the equation (2.35) by

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} \left(\sum_{j=1}^d y_j p_{ij} \right) / \left\{ \sum_{j=1}^d p_{ij} \left(\sum_{h=1}^n \lambda_h^{(k)} p_{hj} \right) \right\} \quad (i = 1, \dots, n), \quad (7.29)$$

under the assumption $\sum_{i=1}^n \sum_{j=1}^d p_{ij} = 1$. They called it the Image Space Reconstruction Algorithm (ISRA), and obtained it heuristically. The operation $\sum_{j=1}^d y_j p_{ij}$ represents a back-projection of the data $\{y_j\}$ from the j th detector. The operation

$$\sum_{j=1}^d p_{ij} \left\{ \sum_{h=1}^n \lambda_h^{(k)} p_{hj} \right\}$$

on the other hand, represents a corresponding back-projection of the current fit for $\{y_j\}$ or the calculated projection. In the ISRA, the ratio between these quantities is multiplicatively used to update the variables. Formal justification of this algorithm was provided later by Titterington (1987) and De Pierro (1989). The motivation for the algorithm, however, is that it requires fewer calculations than the EM algorithm in each iteration. The algorithm does not converge to the MLE based on the Poisson model of Section 2.5, but to a nonnegative least-squares estimate. Convergence properties of the algorithm have been studied and some details and references may be found in De Pierro (1995).

Archer and Titterington (1995) define a more general form of the ISRA suitable for a general class of linear inverse problems with positivity restrictions, mentioned in Section 5.18, analogous to equation (7.29). They then apply it and the EM algorithm to a variety of problems considered in Vardi and Lee (1993). They show that the speeds of the EM algorithm and the ISRA are comparable. However, since the ISRA needs fewer computations in each iteration, it is on the whole more efficient than the EM algorithm.

This Page Intentionally Left Blank

FURTHER APPLICATIONS OF THE EM ALGORITHM

8.1 INTRODUCTION

Since the publication of DLR, the number, variety, and range of applications of the EM algorithm and its extensions have been tremendous. The standard incomplete-data statistical problems such as with missing data, grouped data, censored data, and truncated distributions are being increasingly tackled with EM-type algorithms. Then there are applications of the EM-type of algorithm in a variety of incomplete-data problems arising in standard statistical situations such as linear models, contingency tables and log linear models, random effects models and general variance-components models, and time series and stochastic processes, etc. Special statistical problems such as mixture resolution, factor analysis, survival analysis, and survey sampling have seen increasing use of EM-type of algorithms. These algorithms have also been profitably used in a variety of special applications in engineering, psychometry, econometrics, epidemiology, genetics, astronomy, etc., such as real-time pattern recognition, image processing and reconstruction, autoradiography and medical imaging, especially positron emission tomography (PET), neural networks, and hidden Markov models, and various aspects of signal processing, communication and computing, and AIDS epidemiology and disease prevalence estimation. Applications in many different contexts can be found in the monographs of Little and Rubin (1987, 2002) and McLachlan and Peel (2000a).

Although we have illustrated the methodology of the earlier chapters with many examples, they do not adequately cover the variety and extent of the applications of the EM algorithm and its extensions. Most recent books or review articles on any of these subjects

do include applications of the EM algorithm and its extensions where appropriate. For instance, the mixture resolution problem has attracted a great deal of attention and the books by Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Böhning (1999), Lindsay (1995), McLachlan and Peel (2000a), and Frühwirth-Schnatter (2006), and the review papers by Redner and Walker (1984) and Titterington (1990) contain a great deal of discussion on the EM algorithm in the context of mixture resolution. Some applications like statistical analysis with missing data and analysis of incomplete data from surveys are well covered in Little and Rubin (2002) and Madow, Olkin, and Rubin (1983), respectively. The Little-Rubin book contains EM applications in many different contexts such as regression, linear models, time series, contingency tables, etc. We hope that the exposition of the general theory and methodology in the foregoing chapters will help readers grasp these applications and develop their own when required. We conclude the book with a quick summary of some of the more interesting and topical applications of the EM algorithm.

8.2 HIDDEN MARKOV MODELS

From the large number of illustrations of various aspects of the EM algorithm by the mixture problem, it is evident that the EM algorithm has played a significant part in the solution of the problem of ML estimation of parameters of a mixture. In the mixture framework with observations $\mathbf{w}_1, \dots, \mathbf{w}_n$, the missing-data vector is given by $\mathbf{z} = (z_1^T, \dots, z_n^T)^T$ where, as before, z_j defines the component of the mixture from which the j th data point \mathbf{w}_j arises. We can term this unobserved vector \mathbf{z} as the "hidden variable". In the mixture problems considered up to now, we have been concerned with the fitting of a mixture model, where $\mathbf{y} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T$ denotes an observed random sample. Thus it has sufficed in the complete-data formulation of the problem to take the complete data, $\mathbf{x}_j = (\mathbf{w}_j^T, z_j^T)^T$ ($j = 1, \dots, n$), as a collection of n independent observations.

However, in image analysis, say, where the \mathbf{w}_j refer to intensities measured on n pixels in a scene, the associated component indicator vectors z_j will not be independently distributed because of the spatial correlation between neighboring pixels. In speech recognition applications, the z_j may be unknown serially dependent prototypical spectra on which the observed speech signals \mathbf{w}_j depend. Hence in such cases the sequence or set of hidden values or states z_j cannot be regarded as independent. In these applications, a stationary Markovian model over a finite state space is generally formulated for the distribution of the hidden variable \mathbf{Z} . In one dimension, this Markovian model is a Markov chain (see, for example, Holst and Lindgren, 1991), and in two and higher dimensions a Markov random field (MRF); see Besag (1986, 1989). The conditional distribution of the observed vector \mathbf{Y} is formulated as before to depend only on the value of \mathbf{Z} , the state of the Markov process. As a consequence of the dependent structure of \mathbf{Z} , its probability function does not factor into the product of the marginal probability functions of Z_1, \dots, Z_n in the manner of equation (1.33). However, $\mathbf{W}_1, \dots, \mathbf{W}_n$ are assumed conditionally independent given z_1, \dots, z_n ; that is

$$p(\mathbf{w}_1, \dots, \mathbf{w}_n | z_1, \dots, z_n; \boldsymbol{\theta}) = \prod_{j=1}^n p(\mathbf{w}_j | z_j; \boldsymbol{\theta}),$$

where p is used as a generic symbol for the probability density or the probability mass function and $\boldsymbol{\theta}$ is the parameter vector containing the parameters in these conditional distributions that are known *a priori* to be distinct.

In image-processing applications, the indexing subscript j generally represents pixel-sites and in speech recognition applications, speech frames. In speech recognition applications, \mathbf{W}_j is taken to be a finitely-many-valued vector, representing random functions of the underlying (hidden) prototypical spectra; see Rabiner (1989); Juang and Rabiner (1991). In some applications, the hidden variable Z_j takes values over a continuum representing the gray levels or a similar characteristic of the true image and \mathbf{W}_j is a blurred version of the hidden true image Z_j ; see Qian and Titterington (1991). In some other applications, the hidden variable Z_j may represent some discretized characteristic of the pixel, about which inference is required to be made and \mathbf{W}_j is an observable feature of a pixel statistically related to the hidden characteristic of the pixel.

We first discuss the use of the EM algorithm in a hidden Markov chain. Let the stationary finite Markov chain for Z have state space $\{S_1, \dots, S_g\}$ with transition probability matrix $\mathbf{A} = ((a_{hi}))$, $h, i = 1, \dots, g$. Thus if Z_1, \dots, Z_n is the unknown sequence of states,

$$\text{pr}(Z_{j+1} = S_i | Z_j = S_h) = a_{hi} \quad (h, i = 1, \dots, g).$$

Let the probability distribution of the observed \mathbf{W}_j in state S_i over a common sample space $\{\mathbf{v}_1, \dots, \mathbf{v}_M\}$ be $b_i(\mathbf{v}_m)$, $m = 1, \dots, M$. Thus if $\mathbf{W}_1, \dots, \mathbf{W}_n$ denote the random sequence of observations, then

$$\text{pr}(\mathbf{W}_j = \mathbf{v}_m | Z_j = S_i) = b_i(\mathbf{v}_m) \quad (i = 1, \dots, g; m = 1, \dots, M),$$

not depending on j . Let the initial distribution of the Markov chain be $(\pi_{11}, \dots, \pi_{g1})^T$, ($\sum_{i=1}^g \pi_{i1} = 1$). Thus the parameter vector Ψ for the hidden Markov chain consists of the elements of \mathbf{A} , $\pi_{11}, \dots, \pi_{g-1,1}$, and the $b_i(\mathbf{v}_m)$ for $i = 1, \dots, g$; $m = 1, \dots, M$.

We briefly explain the EM algorithm for this problem, known in the hidden Markov model literature as the Baum-Welch algorithm. Baum and his collaborators formulated this algorithm long before DLR and established convergence properties for this algorithm; see Baum and Petrie (1966), Baum and Eagon (1967), and Baum et al. (1970). We use the following notation:

$$\xi_j(h, i) = \text{pr}_{\Psi}(Z_j = S_h, Z_{j+1} = S_i | \mathbf{Y} = \mathbf{y}),$$

$$\alpha_j(h) = \text{pr}_{\Psi}(\mathbf{W}_1 = \mathbf{w}_1, \dots, \mathbf{W}_j = \mathbf{w}_j, Z_j = S_h),$$

$$\beta_j(h) = \text{pr}_{\Psi}(\mathbf{W}_{j+1} = \mathbf{w}_{j+1}, \mathbf{W}_{j+2} = \mathbf{w}_{j+2}, \dots, \mathbf{W}_n = \mathbf{w}_n | Z_j = S_h),$$

and

$$\gamma_j(h) = \sum_{i=1}^g \xi_j(h, i)$$

for $j = 1, \dots, n-1$; $h, i = 1, \dots, g$. Computing $\xi_j(h, i)$ at current parameter values is essentially the E-step. It can be seen that $\xi_j(h, i)$ can be written with this notation as

$$\xi_j(h, i) = \frac{\alpha_j(h) a_{hi} b_i(w_{j+1}) \beta_{j+1}(i)}{\sum_{h=1}^g \sum_{i=1}^g \alpha_j(h) a_{hi} b_i(w_{j+1}) \beta_{j+1}(i)} \quad (j = 1, \dots, n-1), \quad (8.1)$$

since the numerator is $\text{pr}_{\Psi}(Z_j = S_h, Z_{j+1} = S_i, \mathbf{Y} = \mathbf{y})$ and the denominator is $\text{pr}_{\Psi}(\mathbf{Y} = \mathbf{y})$. To compute this, the values of $\alpha_j(h)$ and $\beta_j(h)$ are to be computed. This is done by forward and backward recursions as follows at the k th iteration:

For $\alpha_j^{(k)}(h)$:

Initialization: $\alpha_1^{(k)}(h) = \pi_h^{(k)} b_h^{(k)}(w_1) \quad (h = 1, \dots, g).$

Induction: $\alpha_{j+1}^{(k)}(i) = [\sum_{h=1}^g \alpha_j^{(k)}(h) a_{hi}^{(k)}] b_i^{(k)}(w_{j+1}) \quad (t = j, \dots, n-1; \quad i = 1, \dots, g).$

Termination: $\text{pr}_{\Psi^{(k)}}(W_1 = w_1, \dots, W_n = w_n) = \sum_{h=1}^g \alpha_n^{(k)}(h).$

For $\beta_j^{(k)}(h)$:

Initialization: $\beta_n^{(k)}(h) = 1 \quad (h = 1, \dots, g).$

Induction: $\beta_j^{(k)}(h) = \sum_{i=1}^g a_{hi}^{(k)} b_i^{(k)}(w_{j+1}) \beta_{j+1}^{(k)}(i)$

for $j = n-1, \dots, 1; \quad i = 1, \dots, g.$

The final computation on the E-step consists of plugging in these values and the current parameter values in the equation (8.1) as follows:

$$\xi_j^{(k)}(h, i) = \frac{\alpha_j^{(k)}(h) a_{hi}^{(k)} b_i^{(k)}(w_{j+1}) \beta_{j+1}^{(k)}(i)}{\sum_{h=1}^g \sum_{i=1}^g \alpha_j^{(k)}(h) a_{hi}^{(k)} b_i^{(k)}(w_{j+1}) \beta_{j+1}^{(k)}(i)} \quad (j = 1, \dots, n-1). \quad (8.2)$$

The M-step consists of finding the updated estimates of the parameters using the following formulas which are a combination of the MLE's for the multinomial parameters and Markov chain transition probabilities based on the number of observed transitions from state h to state i in a finite number of steps observed:

$$\pi_{h1}^{(k+1)} = \gamma_1^{(k)}(h), \quad (8.3)$$

$$a_{hi}^{(k+1)} = \frac{\sum_{j=1}^{n-1} \xi_j^{(k)}(h, i)}{\sum_{j=1}^{n-1} \gamma_j^{(k)}(h)}, \quad (8.4)$$

and

$$b_i^{(k+1)}(\mathbf{v}_m) = \frac{s.t. w_j = \mathbf{v}_m}{\sum_{j=1}^{n-1} \gamma_j^{(k)}(i)}. \quad (8.5)$$

This algorithm is described in some detail in Rabiner (1989). The reader is referred to Leroux and Puterman (1992) for some later work related to this problem. Also, Robert, Celeux, and Diebolt (1993) provide a stochastic Bayesian approach to parameter estimation for a hidden Markov chain.

The EM algorithm for the hidden Markov random field is considerably more difficult; see McLachlan (1992, Chapter 13) and the references therein. Even in the exponential family

case (see Section 1.5.3) the E- and M- steps are difficult to carry out even by numerical methods, except in some very simple cases like a one-parameter case; in some cases they may be implemented by suitable Gibbs sampling algorithms. A variety of practical procedures has been considered in the literature. They are reviewed by Qian and Titterington (1991, 1992), who also suggest a Monte Carlo restoration-estimation algorithm.

In image analysis, the observed data w_j refers to intensities measured on n pixels in a scene, the associated component indicator vectors z_j will not be independently distributed as the intensities between neighboring pixels are spatially correlated. The set of hidden states z_j is viewed as missing data (McLachlan and Peel, 2000a, Chapter 13; van Dyk and Meng, 2001) and a stationary Markovian model over a finite state space is generally formulated for the distribution of the hidden variable Z . An approximation to the E-step, based on a fractional weight version of Besag's iterated conditional modes (ICM) algorithm (Besag, 1986), has been adopted for the segmentation of magnetic resonance images. An alternative approach is a Bayesian one, where the likelihood can be regularized using a prior, resulting in a better-conditioned log likelihood. This can also be interpreted as a penalized likelihood approach. Random field models such as Gibbs priors are often used in this context to capture the local smooth structures of the images. (Geman and Geman, 1984).

Robert et al. (1993) consider a stochastic Bayesian approach to parameter estimation for a hidden Markov chain. Lystig and Hughes (2002) provide a means of implementing a Newton-Raphson approach to obtain parameter estimates and an exact computation of the observed information matrix for hidden Markov models.

Hughes (1997) and Khan (2002) discuss methods for computing the observed information matrix in the hidden Markov model while using the EM algorithm. This will assist in the computation of standard errors for the estimates.

8.3 AIDS EPIDEMIOLOGY

Statistical modeling and statistical methods have played an important role in the understanding of HIV (human immunodeficiency virus) infection and the AIDS (acquired immunodeficiency syndrome) epidemic. Statistical ideas and methods enter naturally in these studies in view of the considerable amount of uncertainty in the data generated for such studies. Further, data from AIDS studies have many new features which need the development of new types of statistical models and methods. For a review of the statistical methods in AIDS studies, see Becker (1992).

Data relating to HIV infection and AIDS epidemics are notoriously incomplete with missing information and delayed reporting and are subject to complex censoring and truncation mechanisms. Furthermore, by the very nature of the disease, there is an unknown and random time-interval between infection and diagnosis. All these aspects result in incompleteness of data and unobservable latent variables, giving rise to EM-type methods. We give a brief summary of the applications of the EM-type algorithms in these problems.

Let us consider the problem of predicting the number of AIDS cases from past data. Let the available data be $\{A_t : t = 1, \dots, T\}$ the number of cases diagnosed as AIDS during a month t (in a certain well-defined population). Let N_t denote the number of individuals infected with HIV in month t . Let f_d be the probability that the duration of infection is d months, defining the probability distribution of incubation period (time between infection and diagnosis) in months. The N_t are unobservable. There are methods to estimate f_d and

hence let us assume that the f_d are known. We have that

$$E(A_t \mid N_1, \dots, N_t) = \sum_{i=1}^t N_i f_{t-i+1}$$

under the assumption of independence of incubation period over individuals. Let $\lambda_i = E(N_i)$ and $\mu_i = E(A_i)$. Then

$$\mu_i = \sum_{i=1}^t \lambda_i f_{t-i+1}.$$

If we use A_t to estimate λ_i , then prediction is done simply by using

$$\mu_{T+r} = \sum_{i=1}^{T+r} \lambda_i f_{T+r-i+1}$$

for the month $T + r$. This is the method of backprojection or backcalculation. Estimation of λ_t is generally done by assuming $\{N_i\}$ to be an inhomogeneous Poisson process. This is an example of an ill-posed inverse problem, much like the problem of PET discussed in Example 2.5 in Section 2.5. Becker, Watson, and Carlin (1991) in fact use an EMS algorithm like that of Silverman et al. (1990) for the PET problem outlined in our Section 5.16. Brookmeyer (1991) and Bacchetti, Segal, and Jewell (1992, 1993) use a penalized likelihood approach, and Pagano, DeGruttola, MaWhinney, and Tu (1992) use ridge regression. In the approach of Bacchetti et al. (1992, 1993), the assumption of $\{N_i\}$ as an inhomogeneous Poisson process is continued. Further, the quantities $a_{ijt} = \text{pr}\{\text{diagnosed at time } j \text{ and reported at time } t \mid \text{infected at time } i\}$ are assumed known and are used in the likelihood. A penalized likelihood with penalizing roughness on the log scale is used. An EM algorithm is developed where the complete data consist of x_{ijt} , denoting the number infected in month i , diagnosed in month j , and reported in month t . Methods based on the penalized EM algorithm (Green, 1990b) are used. The results of the EM algorithm are used as a starting point for fitting a more complicated model.

The foregoing methods need an estimate of the incubation distribution. Bacchetti (1990) discusses estimation of such a distribution (discretized into months) by using data on prospective studies that periodically retest initially seronegative subjects. For those subjects that convert from negative (last being in month L_i , say for subject i) to seropositive (first positive being in month R_i , say), the time of seroconversion t_i is interval-censored. Bacchetti (1990) uses a penalized likelihood method combined with Turnbull's (1974) EM algorithm for general censored and truncated data.

Tu, Meng, and Pagano (1993) estimate the survival distribution after AIDS diagnosis from surveillance data. They use a discrete proportional hazards model, in the presence of unreported deaths, right-truncated sampling of deaths information up to the time of analysis, reporting delays and non-availability of time of death. They extend Turnbull's (1976) approach for regression analysis of truncated and censored data and develop an EM algorithm. They also use the Supplemented EM algorithm for calculating the asymptotic variance of the MLE.

8.4 NEURAL NETWORKS

8.4.1 Introduction

In many important application areas, such as control, pattern recognition, and signal processing, nonlinear adaptive systems are needed to approximate underlying nonlinear mappings through learning from examples. Neural networks (NNs) have been used as such nonlinear adaptive systems, since they can be regarded as universal function approximators of nonlinear functions that can be trained (learned) from examples of input-output data. When the data includes noise, the input-output relation for a neural network is described stochastically in terms of the conditional probability $\text{pr}(\mathbf{u} | \mathbf{w})$ of the output \mathbf{u} given the input \mathbf{w} . Some neural networks (for example, Boltzmann machines) do have explicit probabilistic components in their definition. However, even when a neural network is deterministic, it can be effective to train it as if it were a stochastic network, although it behaves deterministically in the execution model (Amari, 1995b). By working with a stochastic version of a deterministic network, we are able to employ statistical methodology in the learning process of the network (Cheng and Titterington, 1994).

The EM algorithm has been of considerable interest in recent years in the development of algorithms in various application areas of neural networks; see for example, Ma and Ji (1998). Amari (1995b, 1997) has looked at the application of the EM algorithm to neural networks from a geometrical perspective in terms of the information geometry. In related work, Jacobs, Jordan, Nowlan, and Hinton (1991) and Jordan and Xu (1995) have considered the use of the EM algorithm to train mixtures of experts models, which are weighted combinations of subnetworks, while Jordan and Jacobs (1994) and Jacobs, Peng, and Tanner (1997) have considered hierarchical mixtures of experts and the EM algorithm.

For an instance of the use of mixture models in NN's, see Nowlan and Hinton (1993); for the application of NN's in classification, see Ripley (1994); for a treatment of NN's also from a statistical perspective, see Geman, Bienenstock, and Doursat (1992) and Cheng and Titterington (1994). For some other applications of the EM algorithm in neural networks, see Cheeseman, Kelly, Self, Stutz, Taylor, and Freeman (1988) and Nowlan (1991) (in clustering) and Specht (1991) (in density estimation).

Golden (1988) explicitly derives a probability model appropriate for neural network architecture. For example, for a multiple layer associative backpropagation network architecture a suitable model is a conditional multivariate normal density function. The parameters of the probability function are the network's synaptic or connectionist weights. Ma, Ji, and Farmer (1997) and Ma and Ji (1998) attempted a multivariate normal distribution model for feedforward and layered networks. Levin, Tishby, and Solla (1990) considered a layered neural network. Aitkin and Foxall (2003) formulated a statistical approach to the multilayer perceptron. They all used the EM algorithm for maximum likelihood learning of the network weights.

The statistical framework of the EM algorithm allows us to treat the learning process as a maximum likelihood (ML) problem, so standard likelihood-based methodology can be used to train the neural networks in the first instance and to subsequently obtain confidence intervals for a predicted output \mathbf{u} corresponding to an input \mathbf{w} . In the previous chapters, we have seen that there is now a whole battery of EM-related algorithms and more are still being developed. Unfortunately, there exists some misunderstanding about their applications in training a neural network.

In some instances, the conditional expectation of the complete-data log likelihood (the E-step) is effected simply by replacing the random vector of missing data \mathbf{z} by its conditional

expectation. However, this will be valid only if the complete-data log likelihood is linear in \mathbf{z} . Unfortunately it is in general not true and this condition often seems to be neglected in the application of the EM algorithm in the training process of neural networks. Ng and McLachlan (2004a) have attempted to clarify this misconception about the implementation of the EM algorithm in neural networks. They also investigated its application to train multilayer perceptron (MLP) networks and mixture of experts (ME) neural networks in applications to multiclass classification problems. Further, they identified some situations where the application of the EM algorithm for training MLP networks may be of limited value due to complications in performing the E-step.

8.4.2 EM Framework for NNs

In the sequel, we shall assume that the neural network is being used in a multiclass classification context. In the classification context, there are g populations or groups, G_1, \dots, G_g and the problem is to infer the unknown membership of an unclassified entity with a feature vector of p -dimensions. This membership can be defined by a g -dimensional output vector of zero-one indicator variables, where the i th element of the output vector is one or zero, according as the entity does or does not belong to the i th group G_i ($i = 1, \dots, g$). We let

$$(\mathbf{w}_1^T, \mathbf{u}_1^T)^T, \dots, (\mathbf{w}_n^T, \mathbf{u}_n^T)^T \quad (8.6)$$

denote the n classified data points (examples) available for training the neural network, where $\mathbf{w}_j = (w_{1j}, \dots, w_{pj})^T$ is the j th input feature vector and $\mathbf{u}_j = (u_{1j}, \dots, u_{gj})^T$ is the corresponding output vector specifying the group membership of \mathbf{w}_j ($j = 1, \dots, n$). In the training process, the unknown parameters in the neural network, denoted by a vector Ψ , are inferred from the observed training data given by (8.6). We let $\mathbf{w} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T$ and $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_n^T)^T$. In order to estimate Ψ by maximum likelihood, we have to impose a statistical distribution for the observed data, which will allow us to form a log likelihood function. In doing this, we proceed conditionally on the input (feature) vectors in \mathbf{w} .

If the random variables \mathbf{U} and \mathbf{Z} (corresponding to the response \mathbf{u} and the hidden variable \mathbf{z} , respectively) are conditionally independent, then the complete-data log likelihood (conditional on \mathbf{w}) can be expressed as

$$\begin{aligned} \log L_c(\Psi) &\propto \log \text{pr}_{\Psi}(\mathbf{U}, \mathbf{Z} | \mathbf{w}) \\ &= \log \text{pr}_{\Psi}(\mathbf{U} | \mathbf{w}, \mathbf{z}) + \log \text{pr}_{\Psi}(\mathbf{Z} | \mathbf{w}). \end{aligned} \quad (8.7)$$

That is, we need to specify the distribution of the random variable \mathbf{Z} , conditional on \mathbf{w} , and the conditional distribution of \mathbf{U} given \mathbf{w} and \mathbf{z} . On the $(k+1)$ th iteration of the EM algorithm, the E-step computes the Q -function, which is given by

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log L_c(\Psi) | \mathbf{u}, \mathbf{w} \}. \quad (8.8)$$

On the M-step, $\Psi^{(k)}$ is updated by taking $\Psi^{(k+1)}$ to be the value of Ψ that maximizes $Q(\Psi; \Psi^{(k)})$ over all admissible values of Ψ .

As mentioned in 8.4.1, there are instances in the literature of a modified form of the EM algorithm being used unwittingly in that on the E-step, the conditional expectation of the complete-data log likelihood, the Q -function, is effected simply by replacing the random vector \mathbf{z} by its conditional expectation. For example, in Langari, Wang, and Yen (1997) and Wang and Langari (1996), (8.8) is computed by the approximation

$$Q(\Psi; \Psi^{(k)}) \approx \log L_c(\Psi; \mathbf{u}, \tilde{\mathbf{z}}, \mathbf{w}), \quad (8.9)$$

where

$$\tilde{z} = E_{\Psi^{(k)}}\{\mathbf{Z}|\mathbf{u}, \mathbf{w}\}.$$

However, the approximation (8.9) will be valid only in special cases. It is valid if the complete-data log likelihood is linear in \mathbf{z} as in the ME neural networks to be presented in Section 8.4.3, but in general it is not. For instance, it will be seen in the following section that it is a nonlinear function of \mathbf{z} for MLP neural networks, and that it is a quadratic function of \mathbf{z} for the regression models in both the radial basis function (RBF) network of Langari et al. (1997) and the Sugeno-type model of Wang and Langari (1996).

8.4.3 Training Multi-Layer Perceptron Networks

A MLP neural network constructs a decision surface in the data space for discriminating instances with similar features by forming a boundary between them.

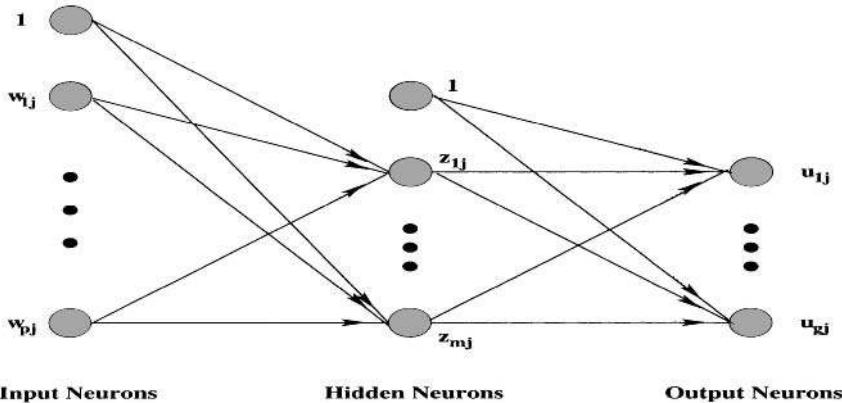


Figure 8.1 A multilayer perceptron neural network.

For a MLP neural network with one hidden layer of m units (Figure 8.1), we can specify a stochastic model of a MLP neural network as follows. Let z_{hj} ($h = 1, \dots, m; j = 1, \dots, n$) be the realization of the zero-one random variable Z_{hj} for which its conditional distribution given \mathbf{w}_j is specified by

$$\text{pr}(Z_{hj} = 1 | \mathbf{w}_j) = \exp(\boldsymbol{\alpha}_h^T \mathbf{w}_j) / \{1 + \exp(\boldsymbol{\alpha}_h^T \mathbf{w}_j)\}, \quad (8.10)$$

where $\boldsymbol{\alpha}_h = (\alpha_{h0}, \alpha_{h1}, \dots, \alpha_{hp})^T$ is the synaptic weight vector of the h th hidden unit. The bias term α_{h0} is included in $\boldsymbol{\alpha}_h$ by adding a constant input $w_{0j} = 1$ for all $j = 1, \dots, n$ so that the input is now $\mathbf{w}_j = (w_{0j}, w_{1j}, \dots, w_{pj})^T$; that is,

$$\boldsymbol{\alpha}_h^T \mathbf{w}_j = \sum_{l=1}^p \alpha_{hl} w_{lj} + \alpha_{h0} = \sum_{l=0}^p \alpha_{hl} w_{lj}.$$

The output of g -dimensional zero-one indicator variables \mathbf{U}_j is distributed according to a multinomial distribution consisting of one draw on g cells with probabilities

$$\text{pr}(U_{ij} = 1 | \mathbf{w}_j, \mathbf{z}_j) = \exp(\boldsymbol{\beta}_i^T \mathbf{z}_j) / \sum_{r=1}^g \exp(\boldsymbol{\beta}_r^T \mathbf{z}_j) \quad (i = 1, \dots, g), \quad (8.11)$$

where $\beta_i = (\beta_{i0}, \beta_{i1}, \dots, \beta_{im})^T$ is the synaptic weight vector of the i th output unit. The bias term β_{i0} is included in β_i by adding a constant hidden unit $z_{0j} = 1$ for all $j = 1, \dots, n$ so that the hidden layer is now $\mathbf{z}_j = (z_{0j}, z_{1j}, \dots, z_{mj})^T$; that is

$$\beta_i^T \mathbf{z}_j = \sum_{h=1}^m \beta_{ih} z_{hj} + \beta_{i0} = \sum_{h=0}^m \beta_{ih} z_{hj}.$$

In the EM framework, the missing data are then given by $\mathbf{z} = (\mathbf{z}_1^T, \dots, \mathbf{z}_n^T)^T$.

The term on the right-hand-side of (8.11) is known as the normalized exponential, or *softmax* function (Bridle, 1990). This function represents a smooth version of the *winner-takes-all* activation model in which the unit with the largest input has output 1 while all other units have output 0. It can be seen from (8.11) that the probabilities are unchanged whenever the same additive constant is added to $\beta_i^T \mathbf{z}_j$ ($i = 1, \dots, g$). For uniqueness, we therefore set $\beta_{gh} = 0$ for $h = 0, 1, \dots, m$. This corresponds to a network with $(g - 1)$ output neurons as, for example, in Figure 8.1. In the case $g = 2$, (8.11) reduces to the logistic transformation. This stochastic specification of the MLP neural network, using the Bernoulli distribution for \mathbf{z}_j and the multinomial distribution for \mathbf{u}_j , has been considered in Amari (1995b). Ma et al. (1997) considered a MLP network in which the output \mathbf{u}_j was linear in \mathbf{z}_j and the stochastic model was specified by assuming the conditional distribution $\text{pr}_{\Psi}(\mathbf{Z} | \mathbf{w})$ to be multivariate normal with known covariance matrix.

In the models specified by (8.10) and (8.11), the vector of all the unknown parameters is given by $\Psi = (\alpha_1^T, \dots, \alpha_m^T, \beta_1^T, \dots, \beta_{g-1}^T)^T$. The ML estimate of Ψ is obtained via the EM algorithm.

Precisely, it follows from (8.10) and (8.11) that

$$\text{pr}(\mathbf{Z} | \mathbf{w}; \Psi) = \prod_{j=1}^n \prod_{h=1}^m v_{hj}^{z_{hj}} (1 - v_{hj})^{(1-z_{hj})},$$

and

$$\text{pr}(\mathbf{U} | \mathbf{w}, \mathbf{z}; \Psi) = \prod_{j=1}^n \prod_{i=1}^g o_{ij}^{u_{ij}},$$

where

$$\begin{aligned} v_{hj} &= \text{pr}(Z_{hj} = 1 | \mathbf{w}_j) \\ &= \frac{\exp(\sum_{l=0}^p \alpha_{hl} w_{lj})}{1 + \exp(\sum_{l=0}^p \alpha_{hl} w_{lj})} \quad (h = 1, \dots, m), \end{aligned} \tag{8.12}$$

and where

$$\begin{aligned} o_{ij} &= \text{pr}(U_{ij} = 1 | \mathbf{w}_j, \mathbf{z}_j) \\ &= \frac{\exp(\sum_{h=0}^m \beta_{ih} z_{hj})}{1 + \sum_{r=1}^{g-1} \exp(\sum_{h=0}^m \beta_{rh} z_{hj})} \quad (i = 1, \dots, g - 1), \end{aligned} \tag{8.13}$$

and

$$\begin{aligned} o_{gj} &= \text{pr}(U_{gj} = 1 | \mathbf{w}_j, \mathbf{z}_j) \\ &= \frac{1}{1 + \sum_{r=1}^{g-1} \exp(\sum_{h=0}^m \beta_{rh} z_{hj})}. \end{aligned} \tag{8.14}$$

From (8.12) and (8.13), the complete-data log likelihood function $\log L_c(\Psi)$ is given by, apart from an additive function not involving Ψ ,

$$\log L_c(\Psi) = \sum_{j=1}^n \left[\sum_{h=1}^m \{z_{hj} \log \frac{v_{hj}}{1-v_{hj}} + \log(1-v_{hj})\} + \sum_{i=1}^g u_{ij} \log o_{ij} \right]. \quad (8.15)$$

It follows on application of the EM algorithm in training the MLP network that on the $(k+1)$ th iteration of the E-step, we need to calculate the Q -function, given by the conditional expectation of $\log L_c(\Psi)$ using the current estimate $\Psi^{(k)}$ for Ψ . From (8.15), the Q -function can be decomposed as

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) &= E_{\Psi^{(k)}} \{ \log L_c(\Psi) \mid \mathbf{u}, \mathbf{w} \} \\ &= \sum_{h=1}^m \sum_{j=1}^n \{ E_{\Psi^{(k)}} (Z_{hj} \mid \mathbf{u}, \mathbf{w}) \log \frac{v_{hj}}{1-v_{hj}} + \log(1-v_{hj}) \} \\ &\quad + \sum_{i=1}^g \sum_{j=1}^n u_{ij} E_{\Psi^{(k)}} (o_{ij} \mid \mathbf{u}, \mathbf{w}) \\ &= Q_\alpha + Q_\beta \end{aligned} \quad (8.16)$$

with respect to the unknown parameters α_h ($h = 1, \dots, m$) and the β_i ($i = 1, \dots, g-1$), respectively.

The first term of the complete-data log likelihood (8.15) is linear in z , so its expectation can be replaced by the expectation of z as in Q_α . The last term of (8.15), however, is nonlinear in z . The decomposition of the Q -function implies that the estimates of α_h and β_i can be updated separately by maximizing Q_α and Q_β , respectively.

We let

$$\tau_{hj}(\mathbf{u}_j, \mathbf{w}_j; \Psi) = E_{\Psi} (Z_{hj} \mid \mathbf{u}_j, \mathbf{w}_j).$$

This conditional expectation of Z_{hj} can be calculated as

$$\begin{aligned} \tau_{hj}(\mathbf{u}_j, \mathbf{w}_j; \Psi) &= E_{\Psi} (Z_{hj} \mid \mathbf{u}_j, \mathbf{w}_j) \\ &= \text{pr}_{\Psi} \{ Z_{hj} = 1 \mid \mathbf{u}_j, \mathbf{w}_j \} \\ &= \text{pr}_{\Psi} \{ Z_{hj} = 1, \mathbf{u}_j \mid \mathbf{w}_j \} / \text{pr}_{\Psi} \{ \mathbf{u}_j \mid \mathbf{w}_j \} \\ &= \sum_{\mathbf{z}_j: z_{hj}=1} \text{pr}_{\Psi} (\mathbf{z}_j, \mathbf{u}_j \mid \mathbf{w}_j) / \sum_{\mathbf{z}_j} \text{pr}_{\Psi} (\mathbf{z}_j, \mathbf{u}_j \mid \mathbf{w}_j) \end{aligned} \quad (8.17)$$

where

$$\text{pr}_{\Psi} (\mathbf{z}_j, \mathbf{u}_j \mid \mathbf{w}_j) = \text{pr}_{\Psi} (\mathbf{z}_j \mid \mathbf{w}_j) \text{pr}_{\Psi} (\mathbf{u}_j \mid \mathbf{z}_j, \mathbf{w}_j). \quad (8.18)$$

From (8.12) and (8.13),

$$\text{pr}_{\Psi} (\mathbf{z}_j \mid \mathbf{w}_j) = \prod_{h=1}^m v_{hj}^{z_{hj}} (1-v_{hj})^{(1-z_{hj})} \quad (8.19)$$

and

$$\text{pr}_{\Psi} (\mathbf{u}_j \mid \mathbf{z}_j, \mathbf{w}_j) = \prod_{i=1}^g o_{ij}^{u_{ij}}. \quad (8.20)$$

On differentiation of Q_α with respect to α_h for $h = 1, \dots, m$, it follows that $\alpha_h^{(k+1)}$ satisfies the equations

$$\sum_{j=1}^n \{\tau_{hj}(\mathbf{u}_j, \mathbf{w}_j; \boldsymbol{\Psi}^{(k)}) - v_{hj}\} \mathbf{w}_j = \mathbf{0} \quad (h = 1, \dots, m). \quad (8.21)$$

On differentiation of Q_β with respect to β_i for $i = 1, \dots, g - 1$, it follows that $\beta_i^{(k+1)}$ satisfies the equations

$$\begin{aligned} & \sum_{j=1}^n u_{ij} \tau_{hj}(\mathbf{u}_j, \mathbf{w}_j; \boldsymbol{\Psi}^{(k)}) \\ & - \sum_{\mathbf{z}_j: z_{hj}=1} o_{ij} \text{pr}_{\boldsymbol{\Psi}^{(k)}}(\mathbf{z}_j, \mathbf{u}_j | \mathbf{w}_j) / \sum_{\mathbf{z}_j} \text{pr}_{\boldsymbol{\Psi}^{(k)}}(\mathbf{z}_j, \mathbf{u}_j | \mathbf{w}_j) = \mathbf{0}. \end{aligned} \quad (8.22)$$

8.4.4 Intractability of the Exact E-Step for MLPs

It can be seen from (8.21) and (8.22) that each E-step of the EM algorithm involves summation over \mathbf{z}_j ($j = 1, \dots, n$). There are 2^m tuples (z_{1j}, \dots, z_{mj}) with $z_{hj} = 0$ or 1 ($h = 1, \dots, m$). Hence, the computational complexity grows exponentially with m . The EM algorithm may provide an efficient training algorithm if the number of hidden units m is small. For example, Lai and Wong (2001) adopted a similar E-step procedure in fitting neural networks to time series data. When m is large, say $m > 10$, a Monte Carlo (MC) approach may be used to implement the E-step.

An alternative variational relaxation approach derived from a mean field approximation has been proposed to circumvent the difficulty of the intractable E-step in training MLP networks (Saul and Jordan, 2000). The basic idea involves approximating the intractable posterior distribution $\text{pr}(\mathbf{Z} | \mathbf{u}, \mathbf{w})$ with a family of factorial distributions, in which the troublesome coupling in the exact EM algorithm in (8.18) and (8.17) are approximated with a factorization assumption for each value of the observed data. A best approximation in the family that is “closest” to the posterior distribution is obtained by minimizing the Kullback-Leibler (KL) divergence between them. By replacing the E-step with an approximate E-step, this variational EM algorithm guarantees global convergence of a lower bound on the log likelihood (Saul and Jordan, 2000).

8.4.5 An Integration of the Methodology Related to EM Training of RBF Networks

For RBF neural networks, the EM algorithm has been used for the unsupervised or supervised modes of the training process (Langari et al., 1997; Streit and Luginbuhl, 1994). In the training of RBF networks, the hidden variable z_{hj} corresponding to the input value \mathbf{w}_j in a RBF neural network has the form

$$z_{hj} = \phi(\mathbf{w}_j; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \quad (h = 1, \dots, m),$$

where $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_m$ denote the radial basis centres and $\boldsymbol{\Sigma}_h$ is a covariance matrix, usually taken to be spherical (that is, $\boldsymbol{\Sigma}_h = \sigma_h^2 I_p$, where I_p is the $p \times p$ identity matrix). The function $\phi(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents the p -variate density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Typically, with RBF neural networks, these centres are found by using a clustering algorithm such as k -means in an unsupervised mode; that is, the clustering algorithm is applied to just the input values \mathbf{w}_j , ignoring their known classification labels. However, attention has been given to using a normal mixture model to find suitable values for the μ_h and the σ_h or the Σ_h before the second stage of finding the weights by conventional neural network training procedures. A detailed description of unsupervised learning of normal mixture models via the EM algorithm can be obtained from McLachlan and Peel (2000a, Chapter 3). Thus, in contrast to the training of MLP networks where a boundary between different groups of instances is sought, training of RBF networks forms clusters in the data space with a centre μ_h for each cluster ($h = 1, \dots, m$). These clusters are then used to classify different groups of data instances. This implies that the training of RBF networks can be substantially faster than the methods used for MLP networks by separately training the basis functions by some unsupervised method and the weights by some linear optimum approaches.

8.4.6 Mixture of Experts

In Section 8.4.4, we have noted some situations where the application of the EM algorithm for training MLP networks may be of limited value due to complications in performing the E-step. In this section, we show that the E-step for training ME neural networks is easy to implement. In ME neural networks (Figure 8.2), there are several modules, referred to as expert networks. These expert networks approximate the distribution of \mathbf{y}_j within each region of the input space. The expert network maps its input \mathbf{w}_j to an output \mathbf{u}_j , with conditional density $f_h(\mathbf{u}_j | \mathbf{w}_j; \boldsymbol{\theta}_h)$, where $\boldsymbol{\theta}_h$ is a vector of unknown parameters for the h th expert network. It is assumed that different experts are appropriate in different regions of the input space. The gating network provides a set of scalar coefficients $\pi_h(\mathbf{w}_j; \boldsymbol{\beta})$ that weight the contributions of the various experts, where $\boldsymbol{\beta}$ is a vector of unknown parameters in the gating network. Therefore, the final output of the ME neural network is a weighted sum of all the output vectors produced by expert networks,

$$f(\mathbf{u}_j | \mathbf{w}_j; \boldsymbol{\Psi}) = \sum_{h=1}^m \pi_h(\mathbf{w}_j; \boldsymbol{\beta}) f_h(\mathbf{u}_j | \mathbf{w}_j; \boldsymbol{\theta}_h), \quad (8.23)$$

where $\boldsymbol{\Psi}$ is the vector of all the unknown parameters. For multiclass classification, the local output density $f_h(\mathbf{u}_j | \mathbf{w}_j; \boldsymbol{\theta}_h)$ is modeled by a multinomial distribution consisting of one draw on g categories.

To apply the EM algorithm to the ME networks, we introduce the indicator variables z_{hj} , where z_{hj} is one or zero according to whether \mathbf{u}_j belongs or does not belong to the h th expert. That is, we let the missing data \mathbf{z} be the vector containing all these indicator variables. The probability that Z_{hj} is one, given the input \mathbf{w}_j , is

$$\pi_h(\mathbf{w}_j; \boldsymbol{\beta}) = \text{pr}\{Z_{hj} = 1 | \mathbf{w}_j\}.$$

The complete-data log likelihood for $\boldsymbol{\Psi}$ is given by

$$\log L_c(\boldsymbol{\Psi}) = \sum_{j=1}^n \sum_{h=1}^m z_{hj} \{\log \pi_h(\mathbf{w}_j; \boldsymbol{\beta}) + \log f_h(\mathbf{u}_j | \mathbf{w}_j; \boldsymbol{\theta}_h)\}. \quad (8.24)$$

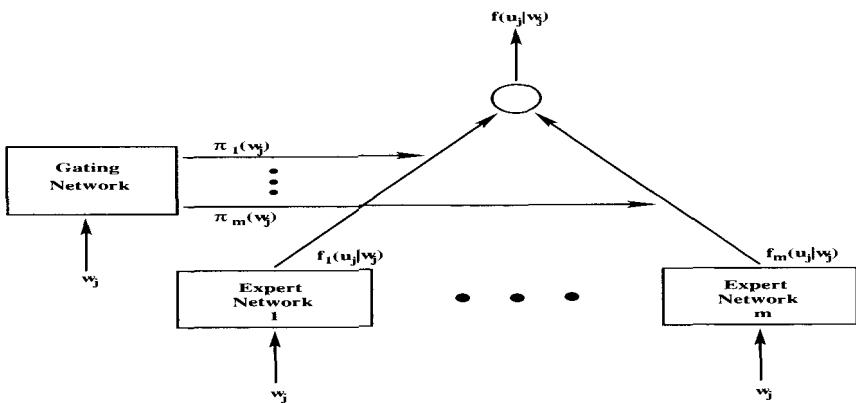


Figure 8.2 Mixture of experts

It follows on application of the EM algorithm in training ME networks that on the $(k+1)$ th iteration, the E-step calculates the Q -function as

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) &= E_{\Psi^{(k)}} \{ \log L_c(\Psi) \mid \mathbf{u}, \mathbf{w} \} \\ &= \sum_{j=1}^n \sum_{h=1}^m E_{\Psi^{(k)}} (Z_{hj} \mid \mathbf{u}, \mathbf{w}) \{ \log \pi_h(\mathbf{w}_j; \beta) + \log f_h(\mathbf{u}_j \mid \mathbf{w}_j; \theta_h) \} \\ &= Q_\beta + Q_\theta. \end{aligned} \quad (8.25)$$

It can be seen that the complete-data log likelihood (8.24) is linear in \mathbf{z} . Thus the E-step just replaces z_{hj} in (8.24) by $\tau_{hj}(\Psi^{(k)}; \mathbf{u}_j \mathbf{w}_j)$, which is its conditional expectation given \mathbf{u}_j and \mathbf{w}_j , formed using $\Psi^{(k)}$ for Ψ . The posterior probability $\tau_{hj}(\Psi; \mathbf{u}_j \mathbf{w}_j)$ can be expressed as

$$\begin{aligned} \tau_{hj}(\Psi; \mathbf{u}_j, \mathbf{w}_j) &= \text{pr}\{Z_{hj} = 1 \mid \mathbf{u}_j, \mathbf{w}_j\} \\ &= \pi_h(\mathbf{w}_j; \beta) f_h(\mathbf{u}_j \mid \mathbf{w}_j; \theta_h) / \sum_{r=1}^m \pi_r(\mathbf{w}_j; \beta) f_r(\mathbf{u}_j \mid \mathbf{w}_j; \theta_r) \end{aligned}$$

for $h = 1, \dots, m$. In addition, the Q -function can be decomposed into two terms with respect to β and θ , respectively.

Hence the M-step consists of two separate maximization problems. The updated estimate of $\beta^{(k+1)}$ is obtained by solving

$$\sum_{j=1}^n \sum_{h=1}^m \tau_{hj}^{(k)} \partial \log \pi_h(\mathbf{w}_j; \beta) / \partial \beta = \mathbf{0}. \quad (8.26)$$

The updated estimate of $\theta_h^{(k+1)}$ is obtained by solving

$$\sum_{j=1}^n \tau_{hj}^{(k)} \partial \log f_h(\mathbf{u}_j \mid \mathbf{w}_j; \theta_h) / \partial \theta_h = \mathbf{0} \quad (8.27)$$

for each h ($h = 1, \dots, m$). Both equations (8.26) and (8.27) require iterative methods. Jordan and Jacobs (1994) proposed an iterative reweighted least squares (IRLS) algorithm for all the generalized linear models used in the ME networks.

The output of the gating network is usually modeled by the multinomial logit (or softmax) function as

$$\pi_h(\mathbf{w}_j; \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}_h^T \mathbf{w}_j)}{1 + \sum_{l=1}^{m-1} \exp(\boldsymbol{\beta}_l^T \mathbf{w}_j)} \quad (h = 1, \dots, m-1), \quad (8.28)$$

where $\pi_m(\mathbf{w}_j; \boldsymbol{\beta}) = 1/(1 + \sum_{l=1}^{m-1} \exp(\boldsymbol{\beta}_l^T \mathbf{w}_j))$. Here $\boldsymbol{\beta}$ contains the elements in $\boldsymbol{\beta}_h$ ($h = 1, \dots, m-1$). Equation (8.26) becomes

$$\sum_{j=1}^n \left(\tau_{hj}^{(k)} - \frac{\exp(\boldsymbol{\beta}_h^T \mathbf{w}_j)}{1 + \sum_{l=1}^{m-1} \exp(\boldsymbol{\beta}_l^T \mathbf{w}_j)} \right) \mathbf{w}_j = 0 \quad (h = 1, \dots, m-1), \quad (8.29)$$

which is a set of nonlinear equations with $(m-1)p$ unknown parameters.

For multiclass classification problems, the h th expert is taken to be the multinomial consisting of one draw on g categories, the local output of the h th expert ($h = 1, \dots, m$) is thus modeled as

$$f_h(\mathbf{u}_j | \mathbf{w}_j; \boldsymbol{\theta}_h) = \prod_{i=1}^{g-1} \left(\frac{\exp(\boldsymbol{\theta}_{hi}^T \mathbf{w}_j)}{1 + \sum_{r=1}^{g-1} \exp(\boldsymbol{\theta}_{hr}^T \mathbf{w}_j)} \right)^{u_{ij}} \left(\frac{1}{1 + \sum_{r=1}^{g-1} \exp(\boldsymbol{\theta}_{hr}^T \mathbf{w}_j)} \right)^{u_{gj}}, \quad (8.30)$$

where $\boldsymbol{\theta}_h$ contains the elements in $\boldsymbol{\theta}_{hi}$ ($i = 1, \dots, g-1$). Equation (8.27) becomes

$$\sum_{j=1}^n \tau_{hj}^{(k)} \left(u_{ij} - \frac{\exp(\boldsymbol{\theta}_{hi}^T \mathbf{w}_j)}{1 + \sum_{r=1}^{g-1} \exp(\boldsymbol{\theta}_{hr}^T \mathbf{w}_j)} \right) \mathbf{w}_j = \mathbf{0} \quad (i = 1, \dots, g-1) \quad (8.31)$$

for $h = 1, \dots, m$, which are m sets of nonlinear equations each with $(g-1)p$ unknown parameters.

It can be seen from (8.29) that the nonlinear equation for the h th expert depends not only on the parameter vector $\boldsymbol{\beta}_h$ but also on other parameter vectors $\boldsymbol{\beta}_l$ ($l = 1, \dots, m-1$). In other words, each parameter vector $\boldsymbol{\beta}_h$ cannot be updated independently. With the IRLS algorithm presented in Jordan and Jacobs (1994), the independence assumption on these parameter vectors was used implicitly and each parameter vector was updated independently and in parallel as

$$\boldsymbol{\beta}_h^{(s+1)} = \boldsymbol{\beta}_h^{(s)} + \gamma_\beta \left(\frac{\partial^2 Q_\beta}{\partial \boldsymbol{\beta}_h \partial \boldsymbol{\beta}_h^T} \right)^{-1} \frac{\partial Q_\alpha}{\partial \boldsymbol{\beta}_h} \quad (h = 1, \dots, m-1), \quad (8.32)$$

where $\gamma_\beta \leq 1$ is the learning rate (Jordan and Xu, 1995). That is, there are $m-1$ sets of nonlinear equations each with p variables instead of a set of nonlinear equations with $(m-1)p$ variables. In Jordan and Jacobs (1994) the iteration (8.32) is referred to as the inner loop of the EM algorithm. This inner loop is terminated when the algorithm is converged or the algorithm is still not converged after some pre-specified number of iterations.

Similarly, each parameter vector $\boldsymbol{\theta}_{hi}$ for $h = 1, \dots, m$ was updated independently as

$$\boldsymbol{\theta}_{hi}^{(s+1)} = \boldsymbol{\theta}_{hi}^{(s)} + \gamma_\theta \left(\frac{\partial^2 Q_\theta}{\partial \boldsymbol{\theta}_{hi} \partial \boldsymbol{\theta}_{hi}^T} \right)^{-1} \frac{\partial Q_\theta}{\partial \boldsymbol{\theta}_{hi}} \quad (i = 1, \dots, g-1). \quad (8.33)$$

where $\gamma_\theta \leq 1$ is the learning rate for \mathbf{w}_{hi} . In the simulation experiment of Ng and McLachlan (2004a) to be discussed in the next section, $\gamma_\alpha = 1$ and γ_θ were set equal to 0.1. They

adopted a smaller learning rate for w_{hi} to ensure better convergence, as u_{ij} in (8.31) is binary zero or one; see the discussion in Ma (1995a, Section 8).

With reference to (8.32) and (8.33), the independence assumption on parameter vectors is equivalent to the adoption of an incomplete Hessian matrix of the Q -function. Chen, Xu, and Chi (1999) proposed a learning algorithm based on the Newton-Raphson method for use in the inner loop of the EM algorithm. In particular, they pointed out that the parameter vectors cannot be updated separately due to the incorrect independence assumption. Rather, they adopted the exact Hessian matrix in the inner loop of the EM algorithm. However, the use of the exact Hessian matrix results in expensive computation during learning. To this end, they proposed a modified algorithm whereby an approximate statistical model called the generalized Bernoulli density is introduced for expert networks in multiclass classification. This approximation simplifies the Newton-Raphson algorithm for multiclass classification in that all off-diagonal block matrices in the Hessian matrix are zero matrices and so the parameter vectors θ_{hi} ($i = 1, \dots, g - 1$) are separable. With this approximation, the learning time is decreased, but the error rate is reported to be increased (Chen et al., 1999).

Ng and McLachlan (2004a) proposed an ECM algorithm for which both parameter vectors β_h and θ_{hi} are separable for $h = 1, \dots, m - 1$ and $i = 1, \dots, g - 1$, respectively. The parameter vector β is partitioned as $(\beta_1^T, \dots, \beta_{m-1}^T)^T$. On the $(k + 1)$ th iteration of the ECM algorithm, the E-step is the same as given above for the EM algorithm. On the M-step, the θ_{hi} are updated in one step. But the updating of β is done over $m - 1$ conditional steps as follows:

- **CM-step 1:** Calculate $\beta_1^{(k+1)}$ by maximizing Q_β with β_l ($l = 2, \dots, m - 1$) fixed at $v_l^{(k)}$.
- **CM-step 2:** Calculate $\beta_2^{(k+1)}$ by maximizing Q_β with β_1 fixed at $\beta_1^{(k+1)}$ and β_l ($l = 3, \dots, m - 1$) fixed at $\beta_l^{(k)}$.
- \vdots
- **CM-step ($m - 1$):** Calculate $\beta_{(m-1)}^{(k+1)}$ by maximizing Q_β with β_l ($l = 1, \dots, m - 2$) fixed at $\beta_l^{(k+1)}$.

As the CM maximizations are over a parameter space of fewer dimensions, they are often simpler and more stable than the corresponding full maximization called for on the M-step of the EM algorithm. More importantly, each CM-step above corresponds to a separable set of the parameters in v_h for $h = 1, \dots, m - 1$, and can be obtained using the IRLS approach. As noted in Chapter 5, the ECM algorithm preserves the appealing convergence properties of the EM algorithm, such as its monotone increasing of likelihood after each iteration.

In more recent work on mixtures of experts, Ng, McLachlan, and Lee (2006) considered an incremental EM-based learning approach in the context of on-line prediction of inpatient length of stay (LOS). Ng and McLachlan (2007) have considered an extension of mixture-of-experts networks for binary classification of correlated data with a hierarchical or clustered structure.

8.4.7 Simulation Experiment

In this section, we report the results of a simulation experiment performed by Ng and McLachlan (2004a) to compare the relative performance of the IRLS algorithm and the ECM algorithm for training the ME networks. Their simulation experiment is similar to that described in Chen et al. (1999), where the IRLS algorithm and the Newton-Raphson algorithm with exact Hessian matrix were compared. In their study, they set both γ_α and γ_θ to be one for the IRLS algorithm and reported that the log likelihood obtained by the IRLS algorithm was oscillatory and unstable and that the error rate was higher.

In the simulated data set, there are three ($g = 3$) groups. As illustrated in Figure 8.3, two small rectangles, an ellipse, and other regions in the large square constitute the three groups denoted G_1 , G_2 , and G_3 , respectively. A training set of ($n = 950$) points are produced by a uniformly distributed random number generator. Among these points, 100, 122, and 728 points belong to G_i ($i = 1, 2, 3$), respectively.

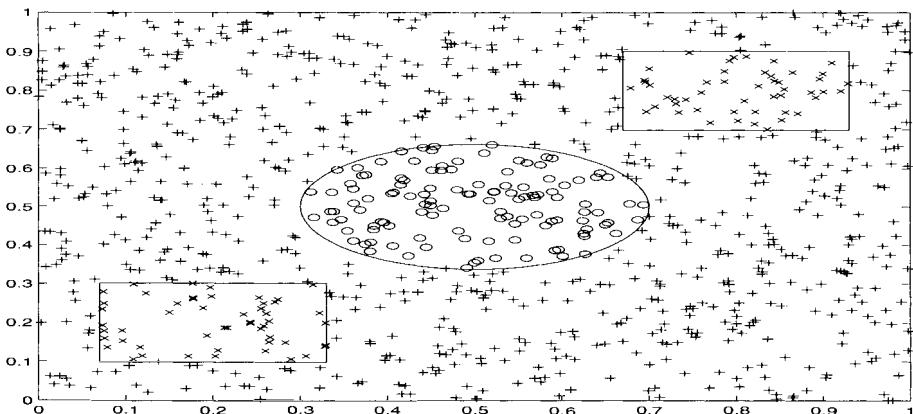


Figure 8.3 Simulated training set; x , o , and $+$ stand for samples belonging to G_1 , G_2 , and G_3 , respectively; From Ng and McLachlan (2004a).

As in Chen et al. (1999), the ME network consists of $m = 12$ experts. For comparative purpose, the learning rates were set by Ng and McLachlan (2004a) to be $\gamma_\alpha = 1$ and $\gamma_\theta = 0.1$ for both IRLS and ECM algorithms. They also ran both algorithms for 60 iterations, where each iteration was composed of a complete E-step and M-step of the EM algorithm. For evaluating the generalization capability, a test set of 2500 points uniformly distributed in the large square was generated. Table 8.1 shows the classification results on both the training and test sets, while in Figure 8.4 , the log likelihood is plotted against the number of iterations.

From Figure 8.4, it can be seen that with both algorithms the log likelihood is increased monotonically after each iteration. Ng and McLachlan (2004a) noted that the unstable behavior of the IRLS algorithm described in Chen et al. (1999) did not occur in this simulation experiment because a learning rate of $\gamma_\theta = 0.1$ was adopted. From Table 8.1, it can be seen that the ECM algorithm outperforms the IRLS algorithm in this simulated experiment, in terms of the misclassified rate for both the training and test sets. Moreover, the ECM algorithm converges to a larger log likelihood value compared to that using the IRLS algorithm (Figure 8.4).

Table 8.1 Simulation Results for the Three-Group Data.

	IRLS Algorithm	ECM Algorithm
Training Set		
No. correctly classified	907	937
No. misclassified	43	13
Test Set		
No. correctly classified	2352	2414
No. misclassified	148	86
Log likelihood	-122.0	-62.4

Source: From Ng and McLachlan (2004a).

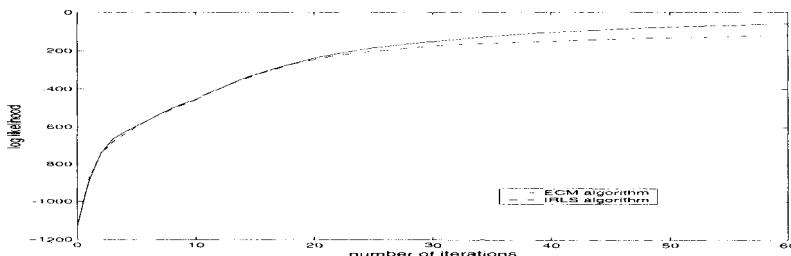


Figure 8.4 Log likelihood versus number of iterations. From Ng and McLachlan (2004a).

8.4.8 Normalized Mixtures of Experts

Xu, Jordan, and Hinton (1995) proposed an alternative form for mixtures of experts, which uses a different parametric form for the gating network. This form is chosen so that the maximization with respect to the parameters of the gating network can be handled analytically. This is achieved by specifying $\pi_h(\mathbf{w}; \boldsymbol{\beta})$ as

$$\pi_h(\mathbf{w}; \boldsymbol{\beta}) = \frac{\alpha_h p_h(\mathbf{w}; \boldsymbol{\xi})}{\sum_{r=1}^m \alpha_r p_r(\mathbf{w}; \boldsymbol{\xi})}. \quad (8.34)$$

The most common example is the Gaussian family for which

$$p_h(\mathbf{w}; \boldsymbol{\xi}) = \phi(\mathbf{w}; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), \quad (8.35)$$

In fitting the mixture of experts model (8.36) with the gating network specified by (8.34), we apply the EM algorithm with the incomplete-data log likelihood formed on the basis of not only \mathbf{u} , but also $\mathbf{w} = (\mathbf{w}_1^T, \dots, \mathbf{w}_j^T)^T$. That is, one learns the parameters of the gating and the expert nets via an asymmetrical representation of the joint density of \mathbf{u} and \mathbf{w} .

Recently, Ng and McLachlan (2005) used normalized Gaussian mixtures of experts to model some data with mixed feature variables; that is, some variables were continuous and some were discrete.

8.4.9 Hierarchical Mixture of Experts

In the previous section, we have concentrated on how the ECM algorithm can be used to train ME networks. The ECM algorithm can also be applied in general to train the hierarchical mixtures of experts (HME) network model of Jordan and Jacobs (1994). In this section, we give the E-and M-steps for an hierarchical mixture of networks.

In a problem of supervised learning in a model of hierarchical mixture of experts, Jordan and Jacobs (1994) and Jordan and Xu (1995) consider the following regression type of model for generating an output \mathbf{u}_j from an input \mathbf{w}_j ,

$$f(\mathbf{u}_j | \mathbf{w}_j; \Psi) = \sum_{h=1}^{g_1} \pi_h(\mathbf{w}_j; \beta_1) \sum_{i=1}^{g_2} \pi_{i \cdot h}(\mathbf{w}_j; \beta_2) p_{hi}(\mathbf{u}_j | \mathbf{w}_j; \theta_{hi}), \quad (8.36)$$

where the $\pi_h(\mathbf{w}_j; \beta_1)$ are multinomial probabilities and the $\pi_{i \cdot h}(\mathbf{w}_j; \beta_2)$ are conditional multinomial probabilities, and Ψ includes both the θ_{hi} parameters (expert network parameters) and the β_1, β_2 parameters (gating network parameters). This is a model for a nested sequence of networks that maps \mathbf{w}_j into \mathbf{u}_j . There is an observed random sample of n observations given by $(\mathbf{u}_j^T, \mathbf{w}_j^T)^T$ ($j = 1, \dots, n$). To tackle the estimation problem, they introduce missing data in terms of the vectors \mathbf{z}_j of zero-one indicator variables with components z_{hj} ($h = 1, \dots, g_1$) for $j = 1, \dots, n$, and the vectors $\mathbf{z}_{\cdot h;j}$ with components $z_{i \cdot h;j}$ ($i = 1, \dots, g_2$) for $h = 1, \dots, g_1; j = 1, \dots, n$. We let $z_{hij} = z_{hj} z_{i \cdot h;j}$. These indicator variables z_{hj} and $z_{i \cdot h;j}$ depend on an input \mathbf{u}_j through their expectations $\pi_h(\mathbf{w}_j; \beta_1)$ and $\pi_{i \cdot h}(\mathbf{w}_j; \beta_2)$, according to the model (8.36). Thus we have \mathbf{z}_j at the first level with the corresponding random variable Z_j having a multinomial distribution consisting of one draw over g categories with probabilities $\{\pi_{hj} = \pi_h(\mathbf{w}_j; \beta_1) : h = 1, \dots, g_1\}$ for each $j = 1, \dots, n$. At the second level, we have conditional on \mathbf{z}_j , the realization $\mathbf{z}_{\cdot h;j}$ with corresponding random variable $Z_{\cdot h;j}$ having a multinomial distribution consisting of one draw over g categories with probabilities $\{\pi_{i \cdot h;j} = \pi_{i \cdot h}(\mathbf{w}_j; \beta_2) : i = 1, \dots, g_2\}$, for each (h, j) for $h = 1, \dots, g_1$ and $j = 1, \dots, n$. We let \mathbf{z} be the vector containing all the labels z_{hj} and $z_{i \cdot h;j}$ ($h, i = 1, \dots, g_1; j = 1, \dots, n$).

If \mathbf{z} were known, then the MLE problem would separate out into regression problems for each expert network and a multiway classification problem for the multinomials. They can be solved independently of each other. A model for the distribution of the j th data point in the complete-data formulation of the problem is

$$p(\mathbf{u}_j, \mathbf{z}_j | \mathbf{w}_j; \Psi) = \prod_{h=1}^{g_1} \prod_{i=1}^{g_2} \{\pi_{hj} \pi_{i \cdot h;j} p_{hi}(\mathbf{u}_j | \mathbf{w}_j; \theta_{hi})\}^{z_{hij}}.$$

Taking the logarithm of this and summing over $j = 1, \dots, n$, we have that the complete-data log likelihood $\log L_c(\Psi)$ is given by

$$\sum_{j=1}^n \sum_{h=1}^{g_1} \sum_{i=1}^{g_2} z_{hij} \{\log \pi_{hj} + \log \pi_{i \cdot h;j} + \log p_{hi}(\mathbf{u}_j | \mathbf{w}_j; \theta_{hi})\}.$$

The EM algorithm then reduces to the following:

E-step. The current conditional expectation $\pi_{hij}^{(k+1)}$ of Z_{hij} is given by

$$\pi_{hij}^{(k+1)} = \frac{\pi_{hj}^{(k)} \pi_{i \cdot h; j}^{(k)} p_{hi}(\mathbf{u}_j \mid \mathbf{w}_j; \boldsymbol{\theta}_{hi}^{(k)})}{\sum_{h=1}^g \sum_{i=1}^g \pi_{hj}^{(k)} \pi_{i \cdot h; j}^{(k)} p_{hi}(\mathbf{u}_j \mid \mathbf{w}_j; \boldsymbol{\theta}_{hi}^{(k)})}.$$

From these joint probabilities $\pi_{hij}^{(k+1)}$, the marginal probabilities $\pi_{hj}^{(k+1)}$ and the conditional probabilities $\pi_{i \cdot h; j}^{(k+1)}$ can be worked out; they correspond to the expected values of Z_{hj} and $Z_{i \cdot h; j}$, respectively.

M-step. The M-step consists of three separate maximization problems:

$$\boldsymbol{\theta}_{hi}^{(k+1)} = \arg \max_{\boldsymbol{\theta}_{hi}} \sum_{j=1}^n \pi_{hij}^{(k+1)} \log p_{hi}(\mathbf{u}_j \mid \mathbf{w}_j; \boldsymbol{\theta}_{hi}),$$

$$\boldsymbol{\beta}_1^{(k+1)} = \arg \max_{\boldsymbol{\beta}_1} \sum_{j=1}^n \sum_{h=1}^{g_1} \pi_{hj}^{(k+1)} \log \pi_h(\mathbf{w}_j; \boldsymbol{\beta}_1),$$

and

$$\boldsymbol{\beta}_2^{(k+1)} = \arg \max_{\boldsymbol{\beta}_2} \sum_{j=1}^n \sum_{h=1}^{g_1} \sum_{i=1}^{g_2} \pi_{hj}^{(k+1)} \pi_{i \cdot h; j}^{(k+1)} \log \pi_{i \cdot h}(\mathbf{w}_j; \boldsymbol{\beta}_2).$$

All these equations require iterative methods and can be solved by IRLS for a generalized linear model.

8.4.10 Boltzmann Machine

We conclude this section on neural networks by looking at the EM algorithm applied to the Boltzmann machine (Ackley, Hinton, and Sejnowski, 1985; Anderson and Titterington, 1998). It is a commonly used form of the NN, behaving according to a discrete-time, stochastic updating rule. The state of the machine at any given time-point is described by a binary n -tuple $\mathbf{w}^T = (w_1, \dots, w_n)$, $w_i \in \{0, 1\}$. At each time-point, i is chosen with equal probability from 1 to n and the value of w_i is ‘updated’ to 1 with probability

$$p_i = \frac{1}{1 + \exp\left\{-\frac{1}{2T} \sum_{j=0}^n a_{ij} w_j\right\}}$$

and to 0 with probability $1 - p_i$, where w_0 is defined to be one. The matrix $\mathbf{A} = ((a_{ij}))$ is generally chosen as a symmetric zero-diagonal matrix. It can be shown that the Boltzmann machine (rather its sequence of states over the time-points) is an ergodic Markov Chain with state space $\{0, 1\}^n$, with a unique stationary distribution

$$p(\mathbf{w}) = b \exp\left[\frac{1}{T} \sum_{i>j} a_{ij} w_i w_j\right], \quad (8.37)$$

being the Boltzmann–Gibbs distribution. In (8.37), b is a normalizing constant. The quantity T is the ‘temperature’ as in Simulated Annealing (Section 7.10.2) and can be used to increase the randomness in the system; it can be seen that as T increases for fixed \mathbf{A} , the stationary distribution tends towards the uniform distribution. In the subsequent discussions, we take T to be 1.

The equation (8.37) parameterizes a very flexible family of distributions. Let us denote the family for a given n by \mathcal{B}_n . Suppose a given probability distribution P over $\{0, 1\}^v$ is to be realized as a stationary distribution of a Boltzmann machine. Suppose that this $P \notin \mathcal{B}_v$. Then it is possible that by constructing a Boltzmann machine with $n (> v)$ variables a better approximation to the given P could be realized than the best from \mathcal{B}_v , as an appropriate v -dimensional marginal distribution of the n -dimensional stationary distribution of the Boltzmann machine. These $n - v$ variables are called hidden variables. In order to study this approximation, we need a criterion. Let us use the Kullback–Leibler divergence criterion between two discrete distributions $p(\mathbf{w})$ and $q(\mathbf{w})$ over the space \mathcal{W} , defined as

$$D(p; q) = \sum_{\mathbf{w} \in \mathcal{W}} p(\mathbf{w}) \log \{p(\mathbf{w})/q(\mathbf{w})\}.$$

Let us denote by \mathcal{D} the class of distributions over $\{0, 1\}^n$ which give rise to the given distribution P as the marginal distribution of the first v variables. Evidently, we would like a Boltzmann machine on n variables which best approximates the given P by the marginal distribution of the first v variables. Then the problem is one of finding a $B \in \mathcal{B}_n$ such that $D(P; B_{(v)})$ is minimized, where $B_{(v)}$ represents the first v -variable marginal of B . This minimization is facilitated by the fact that

$$\min_{B \in \mathcal{B}_n} D(P; B_{(v)}) = \min_{B \in \mathcal{B}_n, P_n \in \mathcal{D}} D(P_n; B),$$

which is proved in Amari, Kurata, and Nagaoka (1992) and Byrne (1992). This minimization can be achieved by an EM algorithm as follows:

Choose an initial distribution $d^{(0)}$ arbitrarily in \mathcal{D} . Then use the following two steps on the k th iteration:

M-step. Find

$$b^{(k+1)} = \arg \min_{b \in \mathcal{B}_n} D(d^{(k)}; b).$$

E-step. Find

$$d^{(k+1)} = \arg \min_{d \in \mathcal{D}} D(d; b^{(k)}).$$

Since the distributions are over $\{0, 1\}^n$, the M-step can be implemented by the Iterative Proportional Fitting (IPF) algorithm. This algorithm is suggested by Csiszár and Tusnády (1984) and is interpreted as an EM algorithm by them as well as by Byrne (1992) and by Neal and Hinton (1998). This algorithm is also called the alternating-minimization algorithm. An interpretation of the alternating-minimization algorithm as an EM algorithm based on the divergence D can be given in more general contexts also. Anderson and Titterington (1995) extend this to a polytomous Boltzmann machine. This approach is also applicable to the general case of NN with hidden units described in the first paragraph of Section 8.4; see Amari (1995a).

8.5 DATA MINING

With the computer revolution, massive data sets of millions of multidimensional observations are now commonplace; see for example, Bishop (2007), Duda, Hart, and Stork (2000), Hand, Mannila, and Smyth (2001), Hastie, Tibshirani, and Friedman (2003), McLachlan

(1992), McLachlan and Peel (2000a), and Ripley (1996), among many other books on or related to data mining. The EM algorithm is being used extensively in data mining as evident in the comparative survey reported recently in Wu et al. (2008). There is an ever increasing demand on speeding up the convergence of the EM algorithm on large databases, as considered in Section 5.19. But at the same time, it is highly desirable if its simplicity and stability can be preserved. In applications where the M-step is computationally simple, for example, in fitting multivariate normal mixtures, the rate of convergence of the EM algorithm depends mainly on the computation time of an E-step as each data point is visited at each E-step. There have been some promising developments on modifications to the EM algorithm for the ML fitting of mixture models to large databases that preserve the simplicity of implementation of the EM in its standard form. For example, McLachlan, Bean, and Peel (2002) have extended the EMMIX (EM-based MIXture) procedure of McLachlan, Peel, Basford, and Adams (1999) to the modeling of high-dimensional data via normal mixtures. Their proposed procedure, which has three steps, is called EMMIX-GENE because it was developed in the context of clustering tissue samples on the basis of the expression levels of thousands of genes (the variables). The first step considers the elimination of variables assessed to have little potential for clustering. On the second step, the retained variables (after appropriate scaling) are clustered into groups essentially using a soft-version of the k -means procedure. Then on the third step, the entities are clustered on the basis of representatives of the groups of variables (metavariables) using mixtures of factor analyzers if the number of metavariables is relatively high. More recently, Ng et al. (2006) have developed a mixture model-based approach (EMMIX-WIRE) to the modeling of high-dimensional data in the case where the correlations between the variables have a known structure that can be formulated through the adoption of a linear mixed model for each cluster. The data can be clustered by fitting mixtures of linear mixed models; see Section 5.10.5.

8.6 BIOINFORMATICS

The literature in the field of bioinformatics is growing at an enormous rate. One area that has attracted much attention in recent times concerns the statistical analysis of high-throughput data, such as microarray gene-expression data. There have been several books written on this and related topics; see, for example, Allison, Page, Beasley, and Edwards (2005), Amaratunga and Cabrera (2003), Biswas, Datta, Fine, and Segal (2008), Ewens and Grant (2005), Gentleman, Carey, Huber, Irizarry, and Dudoit (2005), Keith (2008), Lee (2004), McLachlan, Do, and Ambroise (2004), Parmigiani, Garrett, Irizarry, and Zeger (2003), Simon, Korn, McShane, Radmacher, Wright, and Zhao (2004), Speed (2003), Vannucci, Do, and Müller (2006), and Wit and McClure (2004). As explained in the aforementioned books, the analysis of gene expression microarray data using clustering techniques has an important role to play in the discovery, validation, and understanding of various classes of diseases such as cancer. But the clustering of tumor tissues on the basis of gene expressions is a nonstandard cluster analysis problem since the dimension of each tissue sample is so much greater than the number of tissues. The mixture model-based methods of McLachlan et al. (2002) and Ng et al. (2006) as described briefly in the previous section can be applied to these problems to produce clusters of tissue samples and of gene profiles. These methods rely on the EM algorithm and its variants for their implementation. The EM algorithm is also used elsewhere in bioinformatics, for example, to implement the MEME procedure of Bailey and Elkan (1995) for the discovery of motif patterns in DNA sequences.

REFERENCES

- Aarts, E.H.L. and van Laarhoven, P.J.M. (1989). Simulated annealing: an introduction. *Statistica Neerlandica* **43**, 31–52.
- Achuthan, N.R. and Krishnan, T. (1992). EM algorithm for segregation analysis. *Biometrical Journal* **34**, 971–988.
- Ackley, Hinton, and Sejnowski (1985). A learning algorithm for Boltzmann machines. *Cognitive Science* **9**, 147–169.
- Adamidis, K. and Loukas, S. (1993). ML estimation in the Poisson binomial distribution with grouped data via the EM algorithm. *Journal of Statistical Computation and Simulation* **45**, 33–39.
- Aitkin, M. and Aitkin, I. (1994). Efficient computation of maximum likelihood estimates in mixture distributions, with reference to overdispersion and variance components. In *Proceedings XVIIth International Biometric Conference, Hamilton, Ontario*. Alexandria, VA: Biometric Society, pp. 123–138.
- Aitkin, M. and Aitkin, I. (1996). A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing* **6**, 127–130.
- Aitkin, M. and Foxall, R. (2003). Statistical modelling of artificial neural networks using the multi-layer perceptron. *Statistics and Computing* **13**, 227–239.
- Albert, J.H. (2007). *Bayesian Computation with R (Use R)*. New York: Springer-Verlag.
- Albert, J.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.
- Allison, D.B., Page, G.P., Beasley, T.M., and Edwards, J.W. (Eds.). (2005). *DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments*. Boca Raton, FL: Chapman & Hall/CRC.

- Amaratunga, D. and Cabrera, J. (2003). *Exploration and Analysis of DNA Microarray and Protein Array Data*. Hoboken, NJ: Wiley.
- Amari, S. (1995a). The EM algorithm and information geometry in neural network learning. *Neural Computation* **7**, 13–18.
- Amari, S. (1995b). Information geometry of the EM and em algorithms for neural networks. *IEEE Transactions on Neural Networks* **8**, 1379–1408.
- Amari, S. (1997). Information geometry of neural networks—an overview. In *Mathematics of Neural Networks. Models, Algorithms, and Applications*, S.W. Ellacott, J.C. Mason, and I.J. Anderson (Eds.). Boston: Kluwer, pp. 15–23.
- Amari, S., Kurata, K., and Nagaoka, H. (1992). Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks* **3**, 260–271.
- Amit, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *Journal of Multivariate Analysis* **38**, 82–99.
- Anderson, N.H. and Titterington, D.M. (1995). Beyond the binary Boltzmann machine. *IEEE Transactions on Neural Networks* **6**, 1229–1236.
- Anderson, N.H. and Titterington, D.M. (1998). Boltzmann machines: statistical associations and algorithms for training. In *Neural Network Systems Techniques and Applications*, Vol. 3: *Implementation Techniques*, C.T. Leondes (Ed.). San Diego, CA: Academic Press, pp. 51–89.
- Andrews, D.F. (1974). A robust method for multiple linear regression. *Technometrics* **16**, 523–531.
- Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. New York: Springer-Verlag.
- Andrews, D.F. and Mallows, C.L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B* **36**, 99–102.
- Andrieu, C., Doucet, A., and Robert, C.P. (2004). Computational advances for and from Bayesian analysis. *Statistical Science* **19**, 118–127.
- Archer, G.E.B. and Titterington, D.M. (1995). The iterative image space reconstruction (ISRA) as an alternative to the EM algorithm for solving positive linear inverse problems. *Statistica Sinica* **5**, 77–96.
- Arnold, S.F. (1993). Gibbs sampling. In *Handbook of Statistics*, Vol. 9, C.R. Rao (Ed.). New York: Elsevier, pp. 599–625.
- Arslan, O., Constable, P.D.L., and Kent, J.T. (1993). Domains of convergence for the EM algorithm: a cautionary tale in a location estimation problem. *Communications in Statistics—Theory and Methods* **3**, 103–108.
- Arslan, O., Constable, P.D.L., and Kent, J.T. (1995). Convergence behavior of the EM algorithm for the multivariate t -distribution. *Communications in Statistics—Theory and Methods* **24**, 2981–3000.
- Athreya, K.B., Delampady, M., and Krishnan, T. (2003). Markov chain Monte Carlo methods. (In four parts) *Resonance* **8**, No. 4, 17–26; No. 7, 63–75; No. 10, 9–19; No. 12, 18–32.
- Atkinson, S.E. (1992). The performance of standard and hybrid EM algorithms for ML estimates of the normal mixture model with censoring. *Journal of Statistical Computation and Simulation* **44**, 105–115.
- Avni, Y. and Tananbaum, H. (1986). X-ray properties of optically selected QSO's. *Astrophysics Journal* **305**, 83–99.
- Bacchetti, P. (1990). Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *Journal of the American Statistical Association* **85**, 1002–1008.

- Bacchetti, P., Segal, M.R., and Jewell, N.P. (1992). Uncertainty about the incubation period of AIDS and its impact on backcalculation. In *AIDS Epidemiology: Methodological Issues*, N.P. Jewell, K. Dietz, and V.T. Farewell (Eds.). Boston: Birkhäuser, pp. 61–80.
- Bacchetti, P., Segal, M.R., and Jewell, N.P. (1993). Backcalculation of HIV infection rates (with discussion). *Statistical Science* **8**, 82–119.
- Badawi, R. (1999). *Introduction to PET Physics*. Published on the web at <http://depts.washington.edu/nucmed/IRL/petintro/introsrc/section2.html>
- Bailey, T.L. and Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **21**, 51–80.
- Baker, S.G. (1992). A simple method for computing the observed information matrix when using the EM algorithm. *Journal of Computational and Graphical Statistics* **1**, 63–76.
- Balakrishnan, N. and Kim, J.-A. (2004). EM algorithm for type-II right censored bivariate normal data. In *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*, M.S. Nikulin, N. Balakrishnan, M. Mesbah, and N. Limnios (Eds.). Boston, MA: Birkhäuser, pp. 177–210.
- Balakrishnan, N. and Kim, J.-A. (2005). EM algorithm and optimal censoring schemes for progressively type-II censored bivariate normal data. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability: Methodology and Applications*, N. Balakrishnan, N. Kannan, and H.N. Nagaraja (Eds.). Boston, MA: Birkhäuser, pp. 21–45.
- Bates, D.B. and DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis* **91**, 1–17.
- Bates, D.B. and Pinheiro, J.C. (1998). Computational methods for multilevel modelling. <http://cm.bell-labs.com/cm/ms/departments/sia/project/nlme/CompMulti.pdf>
- Baum, L.E. and Eagon, J.A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society* **73**, 360–363.
- Baum, L.E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite Markov chains. *Annals of Mathematical Statistics* **37**, 1554–1563.
- Baum, L.E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41**, 164–171.
- Beal, M.W.J. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics 7*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West (Eds.). Oxford: Oxford University Press, pp. 453–465.
- Beale, E.M.L. and Little, R.J.A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society B*, **37**, 129–145.
- Becker, N.G. (1992). Statistical challenges of AIDS. *Australian Journal of Statistics* **34**, 129–144.
- Becker, N.G., Watson, L.F., and Carlin, J.B. (1991). A method of non-parametric back-projection and its application to AIDS data. *Statistics in Medicine* **10**, 1527–1542.
- Behboodian, J. (1970). On a mixture of normal distributions. *Biometrika* **57**, 215–217.
- Belin, T.R., and Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association* **90**, 694–707.
- Bentler, P.M. and Tanaka, J.S. (1983). Problems with EM algorithms for ML factor analysis. *Psychometrika* **48**, 247–251.

- Berndt, E.K., Hall, B.H., Hall, R.E., and Hausman, J.A. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* **3/4**, 653–665.
- Bertsimas, D. and Tsitsiklis, J. (1993). Simulated annealing. *Statistical Science* **8**, 10–15.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B* **36**, 192–236.
- Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society B* **48**, 259–302.
- Besag, J. (1989). Towards Bayesian image analysis. *Journal of Applied Statistics* **16**, 395–407.
- Besag, J. (1991). Spatial statistics in the analysis of agricultural field experiments. In *Spatial Statistics and Digital Image Analysis*, Panel on Spatial Statistics and Image Processing (Eds.). Washington, D.C.: National Research Council, National Academy Press, pp. 109–127.
- Besag, J. and Green, P.J. (1993). Spatial statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society B* **55**, 25–37. Discussion: 53–102.
- Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.
- Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**, 1–59.
- Bishop, C.M. (1999). Bayesian PCA. In *Advances in Neural Information Processing Systems, 11*, S.A. Solla, M.S. Kearns, and D.A. Cohn (Eds.). Cambridge, MA: MIT Press, pp. 382–388.
- Bishop, C.M. (2007). *Pattern Recognition and Machine Learning*. New York: Springer-Verlag.
- Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (2007). *Discrete Multivariate Analysis: Theory and Practice*. New York: Springer-Verlag.
- Biswas, A., Datta, S., Fine, J.P., and Segal, M.R. (Eds.). (2008). *Statistical Advances in Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*. Hoboken, NJ: Wiley.
- Blight, B.J.N. (1970). Estimation from a censored sample for the exponential family. *Biometrika* **57**, 389–395.
- Böhning, D. (1993). Construction of reliable maximum likelihood algorithms with application to logistic and Cox regression. In *Handbook of Statistics*, Vol. 9, C.R. Rao (Ed.). Amsterdam: North-Holland, pp. 409–422.
- Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*. New York: Chapman & Hall/CRC.
- Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B.G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics* **46**, 373–388.
- Böhning, D. and Lindsay, B.G. (1988). Monotonicity of quadratic approximation algorithms. *Annals of the Institute of Statistical Mathematics* **40**, 641–663.
- Booth, J.G. and Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo algorithm. *Journal of the Royal Statistical Society B* **61**, 265–285.
- Boyles, R.A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society B* **45**, 47–50.

- Bradley, P.S., Fayyad, U.M., and Reina, C.A. (1998). Scaling EM (expectation-maximization) clustering to large databases. *Technical Report No. MSR-TR-98-35* (revised February, 1999). Seattle: Microsoft Research.
- Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.
- Bridle, J.S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neuro-computing: Algorithms, Architectures and Applications*, F. Fogelman Soulié and J. Hérault (Eds.). Berlin: Springer-Verlag, pp. 227–236.
- Brockwell, P.J. and Davis, R.A. (1996). *Introduction to Time Series Analysis and Forecasting*. New York: Springer-Verlag.
- Broniatowski, M., Celeux, G., and Diebolt, J. (1983). Reconnaissance de densités par un algorithme d'apprentissage probabiliste. In *Data Analysis and Informatics*, Vol. 3. Amsterdam: North-Holland, pp. 359–374.
- Brookmeyer, R. (1991). Reconstruction and future trends of the AIDS epidemic in the United States. *Science* **253**, 37–42.
- Brooks, S.P. and Morgan, B.J.T. (1995). Optimisation using simulated annealing. *The Statistician* **44**, 241–257.
- Brooks, S.P. and Roberts, G.O. (1998). Review of convergence diagnostics. *Statistics and Computing* **8**, 319–335.
- Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*. Second Edition. New York: Wiley.
- Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society B* **22**, 302–306.
- Byrne, C.L. (1993). Iterative image reconstruction algorithms based on cross-entropy minimization. *IEEE Transactions on Image Processing* **2**, 96–103.
- Byrne, W. (1992). Alternating minimization and Boltzmann machine learning. *IEEE Transactions on Neural Networks* **3**, 612–620.
- Cadez, I.V., Smyth, P., McLachlan, G.J., and McLaren, C.E. (2002). Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning* **47**, 7–34.
- Campillo, F. and Le Gland, F. (1989). MLE for partially observed diffusions: Direct maximization vs. the EM algorithm. *Stochastic Processes and their Applications* **33**, 245–274.
- Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B* **57**, 473–484.
- Carlin, J.B. (1987). *Seasonal Analysis of Economic Time Series*. Unpublished Ph.D. Dissertation. Cambridge, MA: Department of Statistics, Harvard University.
- Carter, W.H., Jr. and Myers, R.H. (1973). Maximum likelihood estimation from linear combinations of discrete probability functions. *Journal of the American Statistical Association* **68**, 203–206.
- Casella, G. and Berger, R. (1994). Estimation with selected binomial information or do you really believe that Dave Winfield is batting .471? *Journal of the American Statistical Association* **89**, 1080–1090.
- Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician* **46**, 167–174.

- Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* **2**, 73–82.
- Celeux, G. and Diebolt, J. (1986a). The SEM and EM algorithms for mixtures: numerical and statistical aspects. *Proceedings of the 7th Franco-Belgium Meeting of Statistics*. Bruxelles: Publication des Facultés Universitaires St. Louis.
- Celeux, G. and Diebolt, J. (1986b). L'algorithme SEM: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités. *Revue de Statistique Appliquée* **34**, 35–52.
- Celeux, G. and Diebolt, J. (1990). Une version de type recuit simule de l'algorithme EM. *Comptes Rendus de l'Academie des Sciences. Serie I. Mathematique* **310**, 119–124.
- Celeux, G. and Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastic Reports* **41**, 119–134.
- Ceppellini, R.M., Siniscalco, S., and Smith, C.A.B. (1955). The estimation of gene frequencies in a random-mating population. *Annals of Human Genetics* **20**, 97–115.
- Cerný, V. (1985). Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *Journal of Optimization Theory and Applications* **45**, 41–51.
- Chan, K.S. and Ledolter, J. (1995). Monte Carlo estimation for time series models involving counts. *Journal of the American Statistical Association* **90**, 242–252.
- Chauveau, D. (1995). A stochastic EM algorithm for mixtures with censored data. *Journal of Statistical Planning and Inference* **46**, 1–25.
- Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., and Freeman, D. (1988). AutoClass: a Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, Ann Arbor, Michigan.
- Chen, J., Zhang, D., and Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics* **3**, 347–360.
- Chen, K., Xu, L., and Chi, H. (1999). Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks* **12**, 1229–1252.
- Chen, T.T. (1972). *Mixed-up Frequencies in Contingency Tables*. Ph.D. Dissertation. Chicago: University of Chicago.
- Chen, T.T. and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics* **30**, 629–642.
- Cheng, B. and Titterington, D.M. (1994). Neural networks: a review from a statistical perspective (with discussion). *Statistical Science* **9**, 2–54.
- Chernick, M.R. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers*. Hoboken, NJ: Wiley.
- Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics* **75**, 79–97.
- Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics* **86**, 221–241.
- Chib, S. (2004). Markov chain Monte Carlo technology. In *Handbook of Computational Statistics: Concepts and Methods*, J.E. Gentle, W. Härdle, and Y. Mori (Eds.). New York: Springer-Verlag, pp. 71–102.
- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician* **49**, 327–335.
- Christensen, R. (1990). *Log-linear Models*. New York: Springer-Verlag.

- Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.
- Clayton, D.G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society A* **148**, 82–117.
- Cochran, W.G. and Cox, G. (1957). *Experimental Designs*. New York: Wiley.
- Cohen, M., Dalal, S.R., and Tukey, J.W. (1993). Robust, smoothly heterogeneous variance regression. *Applied Statistics* **42**, 339–353.
- Congdon, P. (2006). *Bayesian Statistical Modelling*. New York: Wiley.
- Cox, D.R. and Hinkley, D. (1974). *Theoretical Statistics*. London: Chapman & Hall.
- Cowell, R.G., Dawid, A.P., Lauritzen, S.L., and Spiegelhalter, D.J. (2003). *Probabilistic Networks and Expert Systems*. Second Edition. New York: Springer-Verlag.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.
- Csiszár, I. and Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions Supplementary Issue No. 1*, 205–237.
- Daube-Witherspoon, M.E. and Muehllehner, G. (1986). An iterative image space reconstruction algorithm suitable for volume ECT. *IEEE Transactions on Medical Imaging* **5**, 61–66.
- Davenport, J.W., Pierce, M.A., and Hathaway, R.J. (1988). A numerical comparison of EM and quasi-Newton type algorithms for computing MLE's for a mixture of normal distributions. *Computer Science and Statistics: Proceedings of the 20th Symposium on the Interface*. Alexandria, VA: American Statistical Association, pp. 410–415.
- Davidon, W.C. (1959). Variable metric methods for minimization. *AEC Research and Development Report ANL-5990*. Argonne, IL: Argonne National Laboratory.
- Day, N.E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463–474.
- De Pierro, A.R. (1989). On some nonlinear iterative relaxation methods in remote sensing. *Matemática Aplicada e Computacional* **8**, 153–166.
- De Pierro, A.R. (1995). A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Transactions on Medical Imaging* **14**, 132–137.
- DeGruttola, V. and Tu, X.M. (1994). Modeling progression of CD-4-lymphocyte count and its relationship to survival time. *Biometrics* **50**, 1003–1014.
- Dellaert, F. (2002). The Expectation–Maximization Algorithm. *Technical Report No. GIT-GVU-02-20*. Atlanta, GA: College of Computing, Georgia Institute of Technology.
- Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: Wiley.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. In *Multivariate Analysis V*, P.R. Krishnaiah (Ed.). Amsterdam: North-Holland, pp. 35–57.
- Dempster, A.P. and Rubin, D.B. (1983). Rounding error in regression: The appropriateness of Sheppard's corrections. *Journal of the Royal Statistical Society B* **45**, 51–59.
- Dennis, J.E. and Schnabel, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

- Didelez, V. and Pigeot, I. (1998). Maximum likelihood estimation in graphical models with missing values. *Biometrika* **85**, 960–966.
- Diebolt, J. and Ip, E.H.S. (1996). Stochastic EM: method and application. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (Eds.). London: Chapman & Hall, pp. 259–273.
- Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B* **56**, 363–375.
- Do, K. and McLachlan, G.J. (1984). Estimation of mixing proportions: a case study. *Applied Statistics* **33**, 134–140.
- Duan, J.-C. and Simonato, J.-G. (1993). Multiplicity of solutions in maximum likelihood factor analysis. *Journal of Statistical Computation and Simulation* **47**, 37–47.
- Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. New York: Wiley.
- Duda, R.O., Hart, P.E., and Stork, G.E. (2000). *Pattern Classification*. Second Edition. New York: Wiley.
- Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. Berkeley, CA: University of California Press, pp. 831–853.
- Efron, B. (1977). Contribution to the discussion of paper by A.P. Dempster, N.M. Laird, and D.B. Rubin. *Journal of the Royal Statistical Society B* **39**, 29.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.
- Efron, B. (1982). Maximum likelihood and decision theory. *Annals of Statistics* **10**, 340–356.
- Efron, B. (1994). Missing data, imputation and the bootstrap (with discussion). *Journal of the American Statistical Association* **89**, 463–479.
- Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika* **65**, 457–487.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.
- Eggermont, P.P.B. and LaRiccia, V.N. (2001). *Maximum Penalized Likelihood Estimation, Vol. I: Density Estimation*. New York: Springer-Verlag.
- Elandt-Johnson, R. (1971). *Probability Models and Statistical Methods in Genetics*. New York: Wiley.
- Elashoff, M. and Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics* **13**, 48–65.
- Enders, C.K. and Peugh, J.L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling* **11**, 1–19.
- Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.
- Everitt, B.S. (1984). Maximum likelihood estimation of the parameters in the mixture of two univariate normal distributions: a comparison of different algorithms. *The Statistician* **33**, 205–215.
- Everitt, B.S. (1987). *Introduction to Optimization Methods and their Application in Statistics*. London: Chapman & Hall.
- Everitt, B.S. (1988). A Monte Carlo investigation of the likelihood-ratio test for number of classes in latent class analysis. *Multivariate Behavioral Research* **23**, 531–538.

- Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*. London: Chapman & Hall.
- Ewens, W.J. and Grant, G.R. (2005) *Statistical Methods in Bioinformatics: An Introduction*. Second Edition. New York: Springer-Verlag.
- Fairclough, D.L., Piermi, W.C., Ridgway, G.J., and Schwertman, N.C. (1992). A Monte Carlo approximation of the smoothing, scoring and EM algorithms for dispersion matrix estimation with incomplete growth curve data. *Journal of Statistical Computation and Simulation* **43**, 77–92.
- Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**, 1041–1046.
- Farewell, V.T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics* **14**, 257–262.
- Fessler, J.A. and Hero, A.O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing* **42**, 2664–2677.
- Fienberg, S.E. (1972). The analysis of multi-way contingency tables. *Biometrics* **28**, 177–202.
- Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A* **222**, 309–368.
- Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.
- Fisher, R.A. (1934). Two new properties of maximum likelihood. *Proceedings of the Royal Society of London A* **144**, 285–307.
- Flury, B. and Zoppé, A. (2000). Exercises in EM. *The American Statistician* **54**, 207–209.
- Fokoué, E. and Titterington, D.M. (2002). Mixtures of factor analyzers. Bayesian estimation and inference by stochastic simulation. *Machine Learning* **50**, 73–94.
- Foulley, J.L. and van Dyk, D.A. (2000). The PX-EM algorithm for fast stable fitting of Henderson's mixed model. *Genetics Selection Evolution* **32**, 143–163.
- Foulley, J.L., Jaffrezic, F., and Robert-Granié, C. (2000). EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data analysis. *Genetics Selection Evolution* **32**, 29–41.
- Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer-Verlag.
- Gamerman, D. and Lopes, H.F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Second Edition. Boca Raton, FL: Chapman & Hall/CRC.
- Ganesalingam, S. and McLachlan, G.J. (1978). The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika* **65**, 658–662.
- Ganesalingam, S. and McLachlan, G.J. (1979). Small sample results for a linear discriminant function estimated from a mixture of normal populations. *Journal of Statistical Computation and Simulation* **9**, 151–158.
- Gelfand, A.E. and Carlin, B.P. (1993). Maximum-likelihood estimation for constrained- or missing-data models. *Canadian Journal of Statistics* **21**, 303–311.
- Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**, 972–985.
- Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.
- Gelfand, A.E., Smith, A.F.M., and Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association* **87**, 523–532.

- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*. Second Edition. Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A. and King, G. (1990). Estimating the electoral consequences of legislative redirecting. *Journal of the American Statistical Association* **85**, 274–282.
- Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457–511.
- Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation* **4**, 1–58.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
- Geman, S. and McClure, D.E. (1985). Bayesian image analysis: an application to single photon emission tomography. *American Statistical Association Proceedings of the Statistical Computing Section*. Alexandria, VA: American Statistical Association, pp. 12–18.
- Geng, Z., Asano, C., Ichimura, M., Tao, F., Wan, K., and Kuroda, M. (1996). Partial imputation method in the EM algorithm. In *Compstat 96*, A. Prat (Ed.). Heidelberg: Physica Verlag, pp. 259–263.
- Gentleman, R., Carey, V., Huber, W., Irizarry, R.A., and Dudoit, S. (Eds.) (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer-Verlag.
- Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–1339.
- Geyer, C.J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7**, 473–511.
- Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte-Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society B* **54**, 657–699.
- Gilks, W.R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (Eds.). Oxford: Oxford University Press, pp. 641–649. Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D., and Kirby, A.J. (1993). Modelling complexity: applications of Gibbs sampling in medicine (with discussion). *Journal of the Royal Statistical Society B* **55**, 39–102.
- Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (Eds.) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348.
- Gill, R.D. (1985). Discussion of the paper by D.Clayton and J.Cuzick. *Journal of the Royal Statistical Society A* **148**, 108–109.
- Gill, R.D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part I) (with discussion). *Scandinavian Journal of Statistics: Theory and Applications* **16**, 97–128.
- Godambe, V.P. and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *International Statistical Review (Revue Internationale de Statistique)* **55**, 231–244.
- Godambe, V.P. and Kale, B.K. (1991). Estimating functions: an overview. In *Estimating Functions*. V.P. Godambe (Ed.). Oxford: Clarendon Press, pp. 3–20.

- Gold, M.S. and Bentler, P.M. (2000). Treatments of missing data: a Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling* **7**, 319–355.
- Golden, R.M. (1988). A unified framework for connectionist systems. *Biological Cybernetics* **59**, 109–120.
- Golub, G.H. and Nash, S.G. (1982). Non-orthogonal analysis of variance using a generalized conjugate gradient algorithm. *Journal of the American Statistical Association* **77**, 109–116.
- Goodman, L.A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology* **79**, 1179–1259.
- Goodnight, J.H. (1979). A tutorial on the sweep operator. *The American Statistician* **33**, 149–158.
- Gordon, N.H. (1990a). Maximum likelihood estimation for mixtures of two Gompertz distributions when censoring occurs. *Communications in Statistics—Simulation and Computation* **19**, 737–747.
- Gordon, N.H. (1990b). Application of the theory of finite mixtures for the estimation of ‘cure’ rates of treated cancer patients. *Statistics in Medicine* **9**, 397–407.
- Graham, J.W. and Hofer, S.M. (2000). Multiple imputation with NORM in a structural equation modeling context. In *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches, and Specific Examples*, T.D. Little, K.U. Schnabel, and J. Baumert (Eds.). Hillsdale, NJ: Lawrence Erlbaum, pp. 201–218.
- Green, P.J. (1990a). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging* **9**, 84–93.
- Green, P.J. (1990b). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society B* **52**, 443–452.
- Grenander, U. (1983). Tutorial in pattern theory. *Technical Report*. Providence, RI: Division of Applied Mathematics, Brown University.
- Haberman, S.J. (1974). Log-linear models for frequency tables derived by indirect observation: maximum likelihood equations. *Annals of Statistics* **2**, 911–924.
- Haberman, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *American Statistical Association Proceedings of the Statistical Computing Section*. Alexandria, VA: American Statistical Association, pp. 45–50.
- Haberman, S.J. (1977). Product models for frequency tables involving indirect observation. *Annals of Statistics* **5**, 1124–1147.
- Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–384.
- Hamilton, J.D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics* **45**, 39–70.
- Hamilton, J.D. (1993). Estimation, inference, and forecasting of time series subject to changes in regime. In *Handbook of Statistics*, Vol. 11: *Econometrics*, G.S. Maddala, C.R. Rao, and H.D. Vinod (Eds.). New York: North-Holland, pp. 231–260.
- Hamilton, J.D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.
- Hand, D.J., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: MIT Press.
- Hartley, H.O. (1958). Maximum likelihood estimation from incomplete data. *Biometrika* **45**, 174–194.

- Hartley, H.O. (1978). Contribution to the discussion of paper by R.E. Quandt and J.B. Ramsey. *Journal of the American Statistical Association* **73**, 738–741.
- Hartley, H.O. and Hocking, R.R. (1971). The analysis of incomplete data (with discussion). *Biometrics* **27**, 783–808.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–340.
- Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics* **8**, 431–444.
- Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association* **64**, 1459–1471.
- Hastie, T., Tibshirani, R., and Friedman, J.H. (2003). *Elements of Statistical Learning*. New York: Springer-Verlag.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
- Hathaway, R.J. (1983). Constrained maximum likelihood estimation for normal mixtures. In *Computer Science and Statistics: The Interface*, J.E. Gentle (Ed.). Amsterdam: North-Holland, pp. 263–267.
- Hathaway, R.J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics* **13**, 795–800.
- Hathaway, R.J. (1986). A constrained EM algorithm for univariate normal mixtures. *Journal of Statistical Computation and Simulation* **23**, 211–230.
- Healy, M.J.R. and Westmacott, M. (1956). Missing values in experiments analyzed on automatic computers. *Applied Statistics* **5**, 203–206.
- Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52**, 271–320.
- Heiser, W.J. (1995). Convergent computing by iterative maximization: theory and applications in multidimensional data analysis. In *Recent Advances in Descriptive Multivariate Analysis*, W.J. Krzanowski (Ed.). Oxford: Clarendon Press, pp. 157–189.
- Heitjan, D.F. (1989). Inference from grouped continuous data: a review (with discussion). *Statistical Science* **4**, 164–183.
- Heitjan, D.F. (1993). Ignorability and coarse data: some biomedical examples. *Biometrics* **49**, 1099–1109.
- Heitjan, D.F. (1994). Ignorability in general incomplete-data models. *Biometrika* **81**, 701–708.
- Heitjan, D.F. and Basu, S. (1996). Distinguishing “Missing at random” and “Missing completely at random”. *The American Statistician* **50**, 207–213.
- Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics* **19**, 2244–2253.
- Hennig, C. (2004). Breakdown points for maximum likelihood-estimators of location-scale mixtures. *Annals of Statistics* **32**, 1313–1340.
- Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**, 185–194.
- Heyde, C.C. and Morton, R. (1996). Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical Society B* **58**, 317–327.
- Hocking, R.R. (2003). *Methods and Applications of Linear Models*. Second Edition. New York: Wiley.
- Holst, U. and Lindgren, G. (1991). Recursive estimation in mixture models with Markov regime. *IEEE Transactions on Information Theory* **37**, 1683–1690.

- Horng, S.C. (1986). *Sublinear Convergence of the EM Algorithm*. Ph.D. Thesis. Los Angeles, CA: University of California.
- Horng, S.C. (1987). Examples of sublinear convergence of the EM algorithm. *American Statistical Association Proceedings of the Statistical Computing Section*. Alexandria, VA: American Statistical Association, pp. 266–271.
- Hosmer, D.W., Jr. (1973a). On MLE of the parameters of a mixture of two normal distributions when the sample size is small. *Communications in Statistics* **1**, 217–227.
- Hosmer, D.W., Jr. (1973b). A comparison of iterative maximum-likelihood estimates of the parameters of a mixture of two normal distributions under three types of sample. *Biometrics* **29**, 761–770.
- Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.
- Hughes, J.P. (1997). Computing the observed information in the hidden Markov model using the EM algorithm. *Statistics & Probability Letters* **32**, 107–114.
- Hunter, D.R. (2003). On the geometry of the EM algorithms. *Technical Report No. 0303*. University Park, PA: Department of Statistics, Penn State University.
- Hunter D.R. (2004). MM algorithms for generalized Bradley-Terry models. *Annals of Statistics* **32**, 386–408.
- Hunter, D.R. and Lange, K. (2000a). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics* **9**, 60–77.
- Hunter, D.R. and Lange, K. (2000b). Rejoinder to discussion of “Optimization transfer using surrogate objective functions.” *Journal of Computational and Graphical Statistics* **9**, 52–59.
- Hunter, D.R. and Lange, K. (2004). A tutorial on MM algorithm. *The American Statistician* **58**, 30–37.
- Ibrahim, J.G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.
- Ikeda, S. (2000). Acceleration of the EM algorithm. *Systems and Computers in Japan* **31**, 10–18.
- Ingrassia, S. (1991). Mixture decomposition via the simulated annealing algorithm. *Applied Stochastic Models and Data Analysis* **7**, 317–325.
- Ingrassia, S. (1992). A comparison between the simulated annealing and the EM algorithms in normal mixture decompositions. *Statistics and Computing* **2**, 203–211.
- Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications* **13**, 151–166.
- Iusem, A.N. (1992). A short convergence proof of the EM algorithm for a specific Poisson model. *Regrage: Revista Brasileira de probabilidade Estatística* **6**, 57–67.
- Jamshidian, M. and Jennrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association* **88**, 221–228.
- Jamshidian, M., and Jennrich, R.I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society B* **62**, 257–270.
- Jacobs, R.A., Jordan, M.I., Nowlan, S.J., and Hinton, G.E. (1991). Adaptive mixtures of local experts. *Neural Computation* **3**, 79–87.
- Jacobs, R.A., Peng, F., and Tanner, M.A. (1997). A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks* **10**, 231–241.
- Jennrich, R.I. and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–820.
- Jolliffe, I.T. (2002). *Principal Component Analysis*. Second Edition. New York: Springer-Verlag.

- Jones, P.N. and McLachlan, G.J. (1990). Algorithm As 254. Maximum likelihood estimation from grouped and truncated data with finite normal mixture models. *Applied Statistics* **39**, 273–282.
- Jones, P.N. and McLachlan, G.J. (1992). Improving the convergence rate of the EM algorithm for a mixture model fitted to grouped truncated data. *Journal of Statistical Computation and Simulation* **43**, 31–44.
- Jordan, M.I. and Jacobs, R.A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6**, 181–214.
- Jordan, M.I. and Xu, L. (1995). Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks* **8**, 1409–1431.
- Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34**, 183–202.
- Jørgensen, B. (1984). The delta algorithm and GLIM. *International Statistical Review (Revue Internationale de Statistique)* **52**, 283–300.
- Juang, B.H. and Rabiner, L.R. (1991). Hidden Markov model for speech recognition. *Technometrics* **33**, 251–272.
- Kay, J. (1994). Statistical models for PET and SPECT data. *Statistical Methods in Medical Research* **3**, 5–21.
- Keith, J.M. (Ed.). (2008). *Bioinformatics*. Totowa, NJ: Humana Press.
- Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. New York: Wiley.
- Kennedy, W.J., Jr. and Gentle, J.E. (1980). *Statistical Computing*. New York: Marcel Dekker.
- Kent, J.T. and Tyler, D.E. (1991). Redescending M-estimates of multivariate location and scatter. *Annals of Statistics* **19**, 2102–2119.
- Kent, J.T., Tyler, D.E., and Vardi, Y. (1994). A curious likelihood identity for the multivariate t -distribution. *Communications in Statistics—Simulation and Computation* **23**, 441–453.
- Khan, R.N. (2002). *Statistical Modelling and Analysis of Ion Channel Data Based on Hidden Markov Models and the EM Algorithm*. Unpublished Ph.D. Thesis. Perth: University of Western Australia.
- Kiefer, N.M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica* **46**, 427–434.
- Kiefer, N.M. (1980). A note on switching regressions and logistic discrimination. *Econometrica* **48**, 1065–1069.
- Kim, D.K., and Taylor, J.M.G. (1995). The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *Journal of the American Statistical Association* **90**, 708–716.
- Kingman, J.F.C. (1993). *Poisson Processes*. Oxford: Oxford University Press.
- Kirkpatrick, S., Gelatt, C.D., Jr. and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.
- Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795–806.
- Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag.
- Kotz, S. and Nadarajah, S. (2004). *Multivariate t-Distributions and Their Applications*. Cambridge: Cambridge University Press.
- Krishnan, T. (1995). EM algorithm in tomography: a review and a bibliography. *Bulletin of Informatics and Cybernetics* **27**, 5–22.

- Krishnan, T. (2004). The EM algorithm. In *Statistical Computing: Existing Methods and Recent Developments*, D.Kundu and A.Basu (Eds.). New Delhi: Narosa, pp. 55–84.
- Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute* **33**, 133–140.
- Lai, T.L. and Wong, S.P.S. (2001). Stochastic neural networks with applications to nonlinear time series. *Journal of the American Statistical Association* **96**, 968–981.
- Laird, N.M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73**, 805–811.
- Laird, N.M. (1982). The computation of estimates of variance components using the EM algorithm. *Journal of Statistical Computation and Simulation* **14**, 295–303.
- Laird, N.M., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association* **82**, 97–105.
- Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.
- Langari, R. Wang, L., and Yen, J. (1997). Radial basis function networks, regression weights, and the expectation-maximization algorithm. *IEEE Transactions on Systems, Man, and Cybernetics A* **27**, 613–623.
- Lange, K. (1990). Convergence of EM image reconstruction algorithm with Gibbs smoothing. *IEEE Transactions on Medical Imaging* **9**, 439–446.
- Lange, K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society B* **57**, 425–437.
- Lange, K. (1995b). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica* **5**, 1–18.
- Lange, K. (1999). *Numerical Analysis for Statisticians*. New York: Springer-Verlag.
- Lange, K. (2004). *Optimization*. New York: Springer-Verlag.
- Lange, K. and Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography* **8**, 306–316.
- Lange, K., Hunter, D.R., and Yang, I. (2000). Optimization using surrogate objective functions. *Journal of Computational and Graphical Statistics* **9**, 1–59.
- Lange, K., Little, R.J.A., and Taylor, J.M.G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881–896.
- Lange, K. and Sinsheimer, S.S. (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* **2**, 175–198.
- Lansky, D., Casella, G., McCulloch, C.E., and Lansky, D. (1992). Convergence and invariance properties of the EM algorithm. *American Statistical Association Proceedings of the Statistical Computing Section*. Alexandria, VA: American Statistical Association, pp. 28–33.
- Larson, M.G. and Dinse, G.E. (1985). A mixture model for the regression analysis of competing risks data. *Applied Statistics* **34**, 201–211.
- Lavielle, M. and Moulines, E. (1997). On a stochastic approximation version of the EM algorithm. *Statistics and Computing* **7**, 229–236.
- Lawley, D.N. and Maxwell, A.E. (1963). *Factor Analysis as a Statistical Method*. London: Butterworths.
- Lawley, D.N. and Maxwell, A.E. (1971). Second Edition. *Factor Analysis as a Statistical Method*. London: Butterworths.

- Lee, A.H., Wang, K., Scott, J.A., Yau, K.K.W., and McLachlan, G.J. (2006). Multilevel zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research* **15**, 47–61.
- Lee, M.-L.T. (2004). *Analysis of Microarray Gene Expression Data*. New York: Springer-Verlag.
- Lehmann, E.L. (1983). *Theory of Point Estimation*. New York: Wiley.
- Lehmann, E.L. and Casella, G. (2003). *Theory of Point Estimation*. New York: Springer-Verlag.
- Leroux, B.G. and Puterman, M.L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48**, 545–558.
- Levin, E., Tishby, N., and Solla, S.A. (1990). A statistical approach to learning and generalization in layered neural network. *Proceedings of the IEEE* **78**, 1568–1574.
- Levine, R.A. and Casella, G. (2001). Implementation of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics* **10**, 422–439.
- Liang, K.-Y. and Zeger, S.L. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science* **10**, 158–173.
- Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry, and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics*, Vol. 5. Hayward, CA: Institute of Mathematical Statistics.
- Lindsay, B.G. and Basak, P.K. (1993). Multivariate normal mixtures: A fast consistent method of moments. *Journal of the American Statistical Association* **88**, 468–476.
- Lindstrom, M.J. and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* **83**, 1014–1022.
- Little, R.J.A. (1983a). Superpopulation models for nonresponse: The ignorable case. In *Incomplete Data in Sample Surveys*, Vol. 2: *Theory and Bibliographies*, W.G. Madow, I. Olkin, and D.B. Rubin (Eds.). New York: Academic Press, pp. 341–382.
- Little, R.J.A. (1983b). Superpopulation models for nonresponse: The nonignorable case. In *Incomplete Data in Sample Surveys*, Vol. 2: *Theory and Bibliographies*, W.G. Madow, I. Olkin, and D.B. Rubin (Eds.). New York: Academic Press, pp. 383–413.
- Little, R.J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics* **37**, 23–38.
- Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.
- Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471–483.
- Little, R.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
- Little, R.J.A. and Rubin, D.B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research* **18**, 292–326.
- Little, R.J.A. and Rubin, D.B. (1990). The analysis of social science data with missing values. In *Modern Methods of Data Analysis*, J. Fox and J.S. Long (Eds.). Newbury Park, CA: Sage Publications, pp. 374–409.
- Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Second Edition. New York: Wiley.
- Liu, C. and Rubin, D.B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633–648.
- Liu, C. and Rubin, D.B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica* **5**, 19–39.

- Liu, C., Rubin, D.B., and Wu, Y.N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* **85**, 755–770.
- Liu, J.S. (1994). The collapsed Gibbs sampler in Bayesian computations with application to a gene regulation problem. *Journal of the American Statistical Association* **89**, 958–966.
- Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag.
- Liu, J.S., Wong, W.H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.
- Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society B* **44**, 226–233.
- Lucy, L.B. (1974). An iterative algorithm for the rectification of observed distributions. *The Astronomical Journal* **79**, 745–754.
- Lystig, T.C. and Hughes, J.P. (2002). Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics* **11**, 678 - 689.
- Ma, S., Ji, C., and Farmer, J. (1997). An efficient EM-based training algorithm for feedforward neural networks. *Neural Networks* **10**, 243–256.
- Ma, S. and Ji, C. (1998). Fast training of recurrent methods based on EM algorithm. *IEEE Transactions on Neural Networks* **9**, 11–26.
- MacEachern, S.N. and Berliner, L.M. (1994). Subsampling the Gibbs Sampler. *The American Statistician* **48**, 188–190.
- Madow, W.G., Olkin, I., and Rubin, D.B. (Eds.). (1983). *Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies*. New York: Academic Press.
- Marin, J.-M. and Robert, C.P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York: Springer-Verlag.
- McColl, J.H., Holmes, A.P., and Ford, I. (1994). Statistical methods in neuroimaging with particular application to emission tomography. *Statistical Methods in Medical Research* **3**, 63–86.
- McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association* **89**, 330–335.
- McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170.
- McCulloch, C.E. (1998). Review of "EM Algorithm and Extensions". *Journal of the American Statistical Association* **93**, 403–404.
- McCulloch, C.E. and Searle, R. (2001). *Generalized, Linear and Mixed Models*. New York: Wiley.
- McGilchrist, C.A. (1994) Estimation in generalized mixed models. *Journal of the Royal Statistical Society B* **56**, 61–69.
- McGilchrist, C.A. and Yau, K.K.W. (1995). The derivation of BLUP, ML, REML, estimation methods for generalised linear mixed models. *Communications in Statistics—Theory and Methods* **24**, 2963–2980.
- McKendrick, A.G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society* **44**, 98–130.
- McLachlan, G.J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association* **70**, 365–369.

- McLachlan, G.J. (1977). Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. *Journal of the American Statistical Association* **72**, 403–406.
- McLachlan, G.J. (1982). The classification and mixture maximum likelihood approaches to cluster analysis. In *Handbook of Statistics*, Vol. 2, P.R. Krishnaiah and L.N. Kanal (Eds.). Amsterdam: North-Holland, pp. 199–208.
- McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**, 318–324.
- McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. (Soft-cover Edition, 2004). New York: Wiley.
- McLachlan, G.J., Adams, P., Ng, S.K., McGiffin, D.C., and Galbraith, A.J. (1994). Fitting mixtures of Gompertz distributions to censored survival data. *Research Report No. 28*. Brisbane: Centre for Statistics, The University of Queensland.
- McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.
- McLachlan, G.J., Basford, K.E., and Green, M. (1993). On inferring the number of components in normal mixture models. *Research Report No. 9*. Brisbane: Centre for Statistics, The University of Queensland.
- McLachlan, G.J., Bean, R.W., and Ben-Tovim Jones, L. (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t distribution. *Computational Statistics and Data Analysis* **51**, 5327–5338.
- McLachlan, G.J., Bean, R.W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.
- McLachlan, G.J., Do, K.-A., and Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. Hoboken, NJ: Wiley.
- McLachlan, G.J. and Jones, P.N. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics* **44**, 571–578.
- McLachlan, G.J. and McGiffin, D.C. (1994). On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research* **3**, 211–226.
- McLachlan, G.J., Ng, S.K., Adams, P., McGiffin, D.C., and Galbraith, A.J. (1997). An algorithm for fitting mixtures of Gompertz distributions to censored survival data. *Journal of Statistical Software* **2**, No. 7.
- McLachlan, G.J. and Peel, D. (2000a). *Finite Mixture Models*. New York: Wiley.
- McLachlan, G.J. and Peel, D. (2000b). Mixtures of factor analyzers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, P. Langley (Ed.). San Francisco, CA: Morgan Kaufmann, pp. 599–606.
- McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. (1999). The EMMIX software for the fitting of mixtures of normal and t -components. *Journal of Statistical Software* **4**, No., 2.
- McLachlan, G.J., Peel, D., and Bean, R.W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis* **41**, 379–388.
- McLachlan, G.J. and Prado, P. (1995). On the fitting of normal mixture models with equal correlation matrices. *Research Report No. 36*. Brisbane: Centre for Statistics, The University of Queensland.
- Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society B* **51**, 127–138.
- Meng, X.-L. (1994). On the rate of convergence of the ECM algorithm. *Annals of Statistics* **22**, 326–339.
- Meng, X.-L. (2007). Thirty years of EM and much more. *Statistica Sinica* **17**, 839–840.

- Meng, X.-L. and Pedlow, S. (1992). EM: a bibliographic review with missing articles. *American Statistical Association Proceedings of the Statistical Computing Section*. Alexandria, VA: American Statistical Association, pp. 24–27.
- Meng, X.-L. and Rubin, D.B. (1989). Obtaining asymptotic variance-covariance matrices for missing-data problems using EM. *American Statistical Association Proceedings of the Statistical Computing Section*. Alexandria, VA: American Statistical Association, pp. 140–144.
- Meng, X.-L. and Rubin, D. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.
- Meng, X.-L. and Rubin, D. (1992). Recent extensions to the EM algorithm (with discussion). In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (Eds.). Oxford: Oxford University Press, pp. 307–320.
- Meng, X.-L. and Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–278.
- Meng, X.-L. and Rubin, D. (1994). On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra and its Applications* **199**, 413–425.
- Meng, X.-L. and van Dyk, D.A. (1997). The EM algorithm—an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society B* **59**, 511–567.
- Mengersen, K., Robert, C.P., and Guihenneuc-Joyaux, C. (1999). MCMC convergence diagnostics: a “reviewww”. In *Bayesian Statistics 6*, J. Berger, J. Bernardo, J.A. Dawid, D. Lindley, and A. Smith Eds.. Oxford: Oxford University Press, pp. 415–440.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.
- Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association* **44**, 335–341.
- Minka, T. (1998). Expectation-Maximization as lower bound maximization. Tutorial published on the web at <http://www-white-media.mit.edu/tp-minka/papers/em.html>.
- Mitchell, T.J. and Turnbull, B.W. (1979). Log-linear models in the analysis of disease prevalence data from survival/sacrifice experiments. *Biometrics* **35**, 221–234.
- Molina, R., Núñez, J., Cortijo, F., and Mateos, J. (2001). Image restoration in astronomy: a Bayesian perspective. *IEEE Signal Processing Magazine* **18**, 11–29.
- Monahan, J.F. (2001). *Numerical Methods of Statistics*. Cambridge: Cambridge University Press.
- Moore, A.W. (1999). Very fast EM-based mixture model clustering using multiresolution kd-trees. In *Advances in Neural Information Processing Systems 11*, M.S. Kearns, S.A. Solla, and D.A. Cohn, (Eds.). Cambridge, MA: MIT Press, pp. 543–549.
- Mosimann, J.E. (1962). On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions. *Biometrika* **49**, 65–82.
- Murray, G.D. (1977). Contribution to the discussion of paper by A.P. Dempster, N.M. Laird, and D.B. Rubin. *Journal of the Royal Statistical Society B* **39**, 27–28.
- Narayanan, A. (1991). Algorithm AS 266: Maximum likelihood estimation of parameters of the Dirichlet distribution. *Applied Statistics* **40**, 365–374.
- Neal, R.N. and Hinton, G.E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, Jordan, M. (Ed). Cambridge, MA: MIT Press, pp. 355–368.

- Nelder, J.A. (1977). Contribution to the discussion of paper by A.P. Dempster, N.M. Laird, and D.B. Rubin. *Journal of the Royal Statistical Society B* **39**, 23–24.
- Nettleton, D. (1999). Convergence properties of the EM algorithm in constrained parameter spaces. *Canadian Journal of Statistics* **27**, 639–648.
- Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics* **8**, 343–366.
- Ng, H.K.T., Chan, P.S., and Balakrishnan, N. (2002). Estimation of parameters from progressively censored data using the EM algorithm. *Computational Statistics and Data Analysis* **39**, 371–386.
- Ng, H.K.T., Chan, P.S., and Balakrishnan, N. (2004). Optimal progressive censoring plans for the Weibull distribution. *Technometrics* **46**, 470–481.
- Ng, S.K., Krishnan, T., and McLachlan, G.J. (2004). The EM algorithm. In *Handbook of Computational Statistics: Concepts and Methods*, J.E. Gentle, W. Härdle, and Y. Mori (Eds.). New York: Springer-Verlag, pp. 135–166.
- Ng, S.K. and McLachlan, G.J. (2003a). On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Statistics and Computing* **13**, 45–55.
- Ng, S.K. and McLachlan, G.J. (2003b). An EM-based semiparametric mixture model approach to the regression analysis of competing-risks data. *Statistics in Medicine* **22**, 1097–111.
- Ng, S.K. and McLachlan, G.J. (2003c). On some variants of the EM algorithm for the fitting of finite mixture models. *Australian Journal of Statistics* **32**, 143–161.
- Ng, S.K. and McLachlan, G.J. (2004a). Using the EM algorithm to train neural networks: misconceptions and a new algorithm for multiclass classification. *IEEE Transactions on Neural Networks* **15**, 738–749.
- Ng, S.K. and McLachlan, G.J. (2004b). Speeding up the EM algorithm for mixture model-based segmentation of magnetic resonance images. *Pattern Recognition* **37**, 1573–1589.
- Ng, S.K. and McLachlan, G.J. (2005). Normalized Gaussian networks with mixed feature data. *Lecture Notes in Artificial Intelligence* **3809**, 879–882.
- Ng, S.K. and McLachlan, G.J. (2007). Extension of mixture-of-experts networks for binary classification of hierarchical data. *Artificial Intelligence in Medicine* **41**, 57–67.
- Ng, S.K., McLachlan, G.J., and Lee, A.H. (2006). An incremental EM-based learning approach for on-line prediction of hospital resource utilization. *Artificial Intelligence in Medicine* **36**, 257–267.
- Ng, S.K., McLachlan, G.J., Wang, K., Ben-Tovim Jones, L., and Ng, S.W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics* **22**, 1745–1752.
- Ng, S.K., McLachlan, G.J., Yau, K.K.W., and Lee, A.H. (2004). A survival mixture model adjusting random hospital effects for analysing ischaemic stroke-specific mortality data. *Statistics in Medicine* **23**, 2729–2744.
- Nielsen, G.G., Gill, R.D., Andersen, P.K., and Sørensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics: Theory and Applications* **19**, 25–43.
- Nowlan, S.J. (1991). Soft competitive algorithm: neural network learning algorithms based on fitting statistical mixtures. *Technical Report No. CMU-CS-91-126*. Pittsburgh, PA: Carnegie-Mellon University.
- Nowlan, S.J. and Hinton, G.E. (1992). Simplifying neural networks by soft weight-sharing. *Neural Computation* **4**, 473–493.

- Nychka, D.W. (1990). Some properties of adding a smoothing step to the EM algorithm. *Statistics & Probability Letters* **9**, 187–193.
- Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society B* **61**, 479–482.
- Orchard, T. and Woodbury, M.A. (1972). A missing information principle: Theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley, CA: University of California Press, pp. 697–715.
- Ostrowski, A.M. (1966). *Solution of Equations and Systems of Equations*. Second Edition. New York: Academic Press.
- Pagano, M., De Gruttola, V., MaWhinney, S., and Tu, X.M. (1992). The HIV epidemic in New York City: statistical methods for projecting AIDS incidence and prevalence. In *AIDS Epidemiology: Methodological Issues*, N.P. Jewell, K. Dietz, and V.T. Farewell (Eds.). Boston, MA: Birkhäuser, pp. 123–142.
- Parmigiani, G., Garrett, E.S., Irizarry, R.A., and Zeger, S.L. (Eds.) (2003). *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer-Verlag.
- Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Clarendon Press.
- Peel, D. and McLachlan, G.J. (2000). Robust mixture modelling using the t distribution. *Statistical Computing* **10**, 335–344.
- Peters, B.C. and Coberly, W.A. (1976). The numerical evaluation of the maximum-likelihood estimate of mixture proportions. *Communications in Statistics—Theory and Methods* **5**, 1127–1135.
- Peters, B.C. and Walker, H.F. (1978). An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal of Applied Mathematics* **35**, 362–378.
- Pettitt, A.N. (1985). Re-weighted least squares estimation with censored and grouped data: an application of the EM algorithm. *Journal of the Royal Statistical Society B* **47**, 253–260.
- Phillips, R.F. (2002). Least absolute deviations estimation via the EM algorithm. *Statistics and Computing* **12**, 281–285.
- Pincus, M. (1968). A closed form solution of certain programming problems. *Operations Research* **16**, 690–694.
- Pincus, M. (1970). A Monte-Carlo method for the approximate solution of certain types of constrained optimization problems. *Operations Research* **18**, 1125–1228.
- Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-Plus*. New York: Springer-Verlag.
- Powell, M.J.D. (1978). A fast algorithm for nonlinearly constrained optimization calculations. In *Lecture Notes in Mathematics No. 630*, G.A. Watson (Ed.). New York: Springer-Verlag, pp. 144–157.
- Propp, J.G. and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.
- Qian, W. and Titterington, D.M. (1991). Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences* **337**, 407–428.
- Qian, W. and Titterington, D.M. (1992). Stochastic relaxations and EM algorithms for Markov random fields. *Journal of Statistical Computation and Simulation* **40**, 55–69.
- Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286.

- Rai, S.N. and Matthews, D.E. (1993). Improving the EM algorithm. *Biometrics* **49**, 587–591.
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. Second Edition. New York: Wiley.
- Ratschek, H. and Rokne, J. (1988). *New Computer Methods for Global Optimization*. New York: Halsted Press.
- Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195–239.
- Richardson, W.H. (1972). Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America* **62**, 55–59.
- Ripley, B.D. (1987). *Stochastic Simulation*. New York: Wiley.
- Ripley, B.D. (1988). *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.
- Ripley, B.D. (1990). Book review of *Simulated Annealing (SA) and Optimization: Modern Algorithms with VLSI, Optimal Design and Missile Defense Applications*, by M.E. Johnson (Ed.). *Journal of Classification* **7**, 287–290.
- Ripley, B.D. (1994). Neural networks and related methods of classification (with discussion). *Journal of the Royal Statistical Society B* **56**, 409–456.
- Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
- Ritter, C. and Tanner, M.A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association* **87**, 861–868.
- Robert, C.P. (1996). Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (Eds.). London: Chapman & Hall, pp. 441–464.
- Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Second Edition. New York: Springer-Verlag.
- Robert, C.P., Celeux, G., and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statistics & Probability Letters* **16**, 77–83.
- Roberts, G.O. and Polson, N.G. (1994). On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society B* **56**, 377–384.
- Roweis, S. (1997). EM algorithms for PCA and SPCA. *Neural Information Processing Systems* **10**, 626–632.
- Rubin, D.B. (1976). Inference with missing data. *Biometrika* **63**, 581–592.
- Rubin, D.B. (1983). Iteratively reweighted least squares. In *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz, N.L. Johnson, and C.B. Read (Eds.). New York: Wiley, pp. 272–275.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Rubin, D.B. (1991). EM and beyond. *Psychometrika* **56**, 241–254.
- Rubin, D.B. and Thayer, D.T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47**, 69–76.
- Rubin, D.B. and Thayer, D.T. (1983). More on EM for factor analysis. *Psychometrika* **48**, 253–257.
- Ruppert, D. and Carroll, R.J. (1980). Trimmed least square estimation in the linear model. *Journal of the American Statistical Association* **75**, 828–838.
- Ruud, P.A. (1991). Extensions of estimation methods using the EM algorithm. *Journal of Econometrics* **49**, 305–341.

- Sahu, S.K. and Roberts, G.O. (1999). On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing* **9**, 55–64.
- Sarkar, A. (1993). On missing data, ignorability and sufficiency. *Technical Report No. 423*. Stanford, CA: Department of Statistics, Stanford University.
- Satten, G.A. and Dutta, S. (2000). A simulate-update algorithm for missing data problems. *Computational Statistics*, **15**, 243–277.
- Saul, L.K. and Jordan, M.I. (2000). Attractor dynamics in feedforward neural networks. *Neural Computation* **12**, 1313–1335.
- Schader, M. and Schmid, F. (1985). Computation of M.L. estimates for the parameters of a negative binomial distribution from grouped data: a comparison of the scoring, Newton-Raphson and E-M algorithms. *Applied Stochastic Models and Data Analysis* **1**, 11–23.
- Schafer, J. (1996). *Analysis by Simulation of Incomplete Multivariate Data*. London: Chapman & Hall.
- Schlossmacher, E.J. (1973). An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association* **68**, 857–859.
- Schmee, J. and Hahn, F.J. (1979). A simple method of regression analysis with censored data. *Technometrics* **21**, 417–434.
- Schneider, T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* **14**, 853–871.
- Segal, M.R., Bacchetti, P., and Jewell, N.P. (1994). Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *Journal of the Royal Statistical Society B* **56**, 345–352.
- Sexton, J. and Swensen, A.R. (2000). ECM algorithms that converge at the rate of EM. *Biometrika* **87**, 651–662.
- Shepp, L.A. and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging* **1**, 113–122.
- Shih, W.J. and Weisberg, S. (1986). Assessing influence in multiple linear regression with incomplete data. *Technometrics* **28**, 231–239.
- Shumway, R.H. and Stoffer, D.S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* **3**, 253–264.
- Shumway, R.H. and Stoffer, D.S. (2000). *Time Series Analysis and its Applications*. New York: Springer-Verlag.
- Silverman, B.W., Jones, M.C., Wilson, J.D., and Nychka, D.W. (1990). A smoothed EM approach to indirect estimation – A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *Journal of the Royal Statistical Society B* **52**, 271–324.
- Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W., and Zhao, Y. (Eds.). (2004). *Design and Analysis of DNA Microarray Investigations*. New York: Springer-Verlag.
- Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society B* **55**, 3–23. Discussion: 53–102.
- Smith, C.A.B. (1957). Counting methods in genetical statistics. *Annals of Human Genetics* **21**, 254–276.
- Smith, C.A.B. (1977). Contribution to the discussion of the paper by A.P. Dempster, N.M. Laird, and D.B. Rubin. *Journal of the Royal Statistical Society B* **39**, 24–25.

- Sobel, E. and Lange, K. (1994). Metropolis sampling in pedigree analysis. *Statistical Methods in Medical Research* **2**, 263–282.
- Sørenson, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. New York: Springer-Verlag.
- Specht, D.F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks* **2**, 568–576.
- Speed, T.P. (Ed.). (2003). *Statistical Analysis of Gene Expression Microarray Data*. Boca Raton, FL: Chapman & Hall/CRC.
- Steele, B.M. (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrics* **52**, 1295–1310.
- Streit, R.L. and Luginbuhl, T.E. (1994). Maximum likelihood training of probabilistic neural networks. *IEEE Transactions on Neural Networks* **5**, 764–783.
- Stuart, A. and Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics*, Vol. 2: *Classical Inference and Relationship*. Sixth Edition. London: Charles Griffin.
- Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics: Theory and Applications* **1**, 49–58.
- Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in Statistics—Simulation and Computation* **5**, 55–64.
- Svensén, M. and Bishop, C.M. (2005). Robust Bayesian mixture modelling. *Neurocomputing* **64**, 234–252.
- Tan, W.Y. and Chang, W.C. (1972). Convolution approach to genetic analysis of quantitative characters of self-fertilized populations. *Biometrics* **28**, 1073–1090.
- Tanner, M.A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. Lecture Notes in Statistics, Vol. 67. New York: Springer-Verlag.
- Tanner, M.A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Second Edition. New York: Springer-Verlag.
- Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.
- Thiesson, B. (1995). Accelerated quantification of Bayesian networks with incomplete data. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, U. Fayyad and R. Uthuruswamy (Eds.). Menlo Park, CA: AAAI Press, pp. 306–311.
- Thiesson, B. (1997). Score and information for recursive exponential models with incomplete data. In *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, D. Geiger and P. Shenoy (Eds.). San Francisco, CA: Morgan Kaufmann, pp. 453–463.
- Thisted, R.A. (1988). *Elements of Statistical Computing: Numerical Computation*. London: Chapman & Hall.
- Thompson, E.A. (1975). *Human Evolutionary Trees*. Cambridge: Cambridge University Press.
- Thompson, E.A. (1977). Contribution to the discussion of the paper by A.P. Dempster, N.M. Laird, and D.B. Rubin. *Journal of the Royal Statistical Society B* **39**, 33–34.
- Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**, 1701–1762.

- Tipping, M.E. and Bishop, C.M. (1997). Mixtures of probabilistic principal component analysers. *Technical Report No. NCRG/97/003*. Aston University, Birmingham: Neural Computing Research Group.
- Tipping, M.E. and Bishop, C.M. (1999a). Probabilistic principal component analysis. *Journal of the Royal Statistical Society B* **61**, 611–622.
- Tipping, M.E. and Bishop, C.M. (1999b). Mixtures of probabilistic principal component analyzers. *Neural Computation* **11**, 443–482.
- Titterington, D.M. (1984). Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society B* **46**, 257–267.
- Titterington, D.M. (1987). On the iterative image space reconstruction algorithm for ECT. *IEEE Transactions on Medical Imaging* **6**, 52–56.
- Titterington, D.M. (1990). Some recent research in the analysis of mixture distributions. *Statistics* **21**, 619–641.
- Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.
- Tu, X.M., Meng, X.-L., and Pagano, M. (1993). The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association* **88**, 26–36.
- Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association* **69**, 169–173.
- Turnbull, B.W. (1976). The empirical distribution with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society B* **38**, 290–295.
- Turnbull, B.W. and Mitchell, T.J. (1978). Exploratory analysis of disease prevalence data from survival/sacrifice experiments. *Biometrics* **34**, 555–570.
- Turnbull, B.W. and Mitchell, T.J. (1984). Nonparametric estimation of the distribution of time to onset for specific diseases in survival/sacrifice experiments. *Biometrics* **40**, 41–50.
- Vaida, F. and Meng, X.-L. (2005). Two-slice EM algorithms for fitting generalized linear mixed models with binary response. *Statistical Modelling* **5**, 229–242.
- Vaida, F., Meng, X.-L., and Xu, R. (2004). Mixed effects models and the EM algorithm. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* A. Gelman and X.-L. Meng (Eds.). New York: Wiley, pp. 253–264.
- van Dyk, D.A. (2000). Fitting mixed-effects models using efficient EM-type algorithms. *Journal of Computational and Graphical Statistics* **9**, 78–98.
- van Dyk, D.A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* **10**, 101–111.
- van Dyk, D.A. and Tang, R. (2003). The one-step-late PXEM algorithm. *Statistics and Computing* **13**, 137–152.
- van Laarhoven, P.J.M. and Aarts, E.H.L. (1987). *Simulated Annealing: Theory and Applications*. Dordrecht: D.Reidel Publishing Company.
- Vannucci, M., Do, K.-A., and Müller, P. (Eds.). (2006). *Bayesian Inference for Gene Expression and Proteomics*. Cambridge: Cambridge University Press.
- Vardi, Y. and Lee, D. (1993). From image deblurring to optimal investments: maximum likelihood solutions to positive linear inverse problems (with discussion). *Journal of the Royal Statistical Society B* **55**, 569–612.
- Vardi, Y., Shepp, L.A., and Kaufman, L. (1985). A statistical model for positron emission tomography (with discussion). *Journal of the American Statistical Association* **80**, 8–37.

- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.
- Wang, C.Y., Huang, Y., Chao, E.C., and Jeffcoat, M.K.. (2007). Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. *Biometrics (OnlineEarly Articles)*. doi:10.1111/j.1541-0420.2007.00839.x
- Wang, L. and Langari, R. (1996). Sugeno model, fuzzy discretization, and the EM algorithm. *Fuzzy Systems* **82**, 279–288.
- Wang, B. and Titterington, D.M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis* **1**, 625–650.
- Watson, M. and Engle, R.F. (1983). Alternative algorithms for the estimation of dynamic factor, MIMIC and varying coefficient regression models. *Journal of Econometrics* **23**, 385–400.
- Wei, G.C.G. and Tanner, M.A. (1990a). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704.
- Wei, G.C.G. and Tanner, M.A. (1990b). Posterior computations for censored regression data. *Journal of the American Statistical Association* **85**, 829–839.
- Weiss, G.H. (1989). Simulated Annealing. In *Encyclopedia of Statistical Sciences*, Supplementary Volume, S. Kotz, N.L. Johnson, and C.B. Read (Eds.). New York: Wiley, pp. 144–146.
- Wikipedia Contributors (2007). *Positron Emission Tomography* [Internet]. Wikipedia, The Free Encyclopedia. Available from: <http://en.wikipedia.org/w/index.php?title=Positron%20emission%20tomography&oldid=163436515>.
- Wit, E. and McClure, J. (2004). *Statistics for Microarrays: Design, Analysis and Inference*. Hoboken, NJ: Wiley.
- Wolfe, J.H. (1967). NORMIX: Computational methods for estimating the parameters of multivariate normal mixtures of distributions. *Research Memo. SRM 68-2*. San Diego, CA: U.S. Naval Personnel Research Activity.
- Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* **5**, 329–350.
- Wolynetz, M.S. (1979a). Algorithm AS 138: Maximum likelihood estimation from confined and censored normal data. *Applied Statistics* **28**, 185–195.
- Wolynetz, M.S. (1979b). Algorithm AS 139: Maximum likelihood estimation in a linear model from confined and censored normal data. *Applied Statistics* **28**, 195–206.
- Wolynetz, M.S. (1980). A remark on Algorithm 138: maximum likelihood estimation from confined and censored normal data. *Applied Statistics* **29**, 228.
- Wright, K. and Kennedy, W.J., Jr. (2000). An interval analysis approach to the EM algorithm. *Journal of Computational and Graphical Statistics* **9**, 303–318.
- Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103.
- Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, S.K., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., and Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*. To appear.
- Wulfsohn, M.S., and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measures with error. *Biometrics* **53**, 330–339.

- Xiang, L., Lee, A.H., Yau, K.K.W., and McLachlan, G.J. (2006). A score test for zero-inflation in correlated count data. *Statistics in Medicine* **25**, 1660–1670.
- Xiang, L., Lee, A.H., Yau, K.K.W., and McLachlan, G.J. (2007). A score test for overdispersion in zero-inflated Poisson mixed regression model. *Statistics in Medicine* **26**, 1608–1622.
- Xu, L. and Jordan, M.I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation* **8**, 129–151.
- Xu, L., Jordan, M.I., and Hinton, G.E. (1995). An alternative model for mixtures of experts. In *Advances in Neural Information Processing Systems*, J.D. Cowan, G. Tesauro, and J. Alspector (Eds.). Cambridge, MA: MIT Press, pp. 663–640.
- Yates, F.Y. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture* **1**, 129–142.
- Yau, K.K.W., Lee, A.H., and Ng, S.K. (2003). Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics and Data Analysis* **41**, 359–366.
- Yuille, A.L., Stolorz, P., and Utans, J. (1994). Statistical physics, mixtures of distributions, and the EM algorithm. *Neural Computation* **6**, 334–340.
- Zangwill, W.I. (1969). *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall.
- Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86.
- Zhao, J. and Jiang, Q. (2006). Probabilistic PCA for t distributions. *Neurocomputing* **69**, 2217–2226.

This Page Intentionally Left Blank

AUTHOR INDEX

- Aarts, E.H.L., 285, 311, 335
Achuthan, N.R., 69, 311
Ackley, D.H., 308, 311
Adamidis, K., 66, 311
Adams, P., 172, 173, 310, 328
Aitkin, I., 32, 146, 147, 286, 311
Aitkin, M., 32, 146, 147, 286, 295, 311
Albert, J.H., 238, 250, 311
Allison, D.B., 310, 311
Amaratunga, D., 310, 312
Amari, S., 32, 295, 298, 309, 312
Ambroise, C., 310, 328
Amit, Y., 258, 312
Andersen, P.K., 173, 330
Anderson, N.H., 308, 309, 312
Andrews, D.F., 32, 61, 62, 168, 312
Andrieu, C., 254, 312
Archer, G.E.B., 287, 312
Arnold, S.F., 242, 312
Arslan, O., 88, 90, 200, 312
Asano, C., 36, 320
Athreya, K.B., xx, 233, 238, 312
Atkinson, S.E., 146, 312
Avni, Y., 32, 312
- Bacchetti, P., 34, 214, 215, 294, 312, 313, 333
Badawi, R., 55, 313
Bailey, T.L., 310, 313
Baker, S.G., 34, 124, 131, 135, 313
Balakrishnan, N., 173, 313, 330
Basak, P.K., 286, 326
Basford, K.E., 4, 13, 17, 31, 32, 63, 65, 115,
 168, 290, 310, 328
Basu, S., 267, 322
Bates, D.B., 313
Bates, D.M., 142, 188, 191, 326, 331
Baum, L.E., 30, 82, 256, 291, 313
Beal, M.W.J., 276, 277, 313
Beale, E.M.L., 31, 313
Bean, R.W., 195, 207, 208, 310, 328
Beasley, T.M., 310, 311
Becker, N.G., 293, 294, 313
Behboodian, J., 33, 313
Belin, T.R., 125, 313
Ben-Tovim Jones, L., 191, 208, 328, 330
Bentler, P.M., 196, 267, 313, 321
Berger, R., 256, 315
Berliner, L.M., 237, 327
Berndt, E.K., 115, 314
Bertsimas, D., 285, 314
- Besag, J., 32, 185, 238, 239, 242, 290, 293,
 314
Best, N.G., 187, 190, 238, 242, 320
Bienenstock, E., 295, 320
Bishop, C.M., 196, 197, 207, 309, 314, 334,
 335
Bishop, Y.M.M., 162, 314
Biswas, A., 310, 314
Blight, B.J.N., 30, 33, 314
Böhning, D., 6, 142, 144, 286, 290, 314
Booth, J.G., 192, 222–225, 249, 314
Boyles, R.A., 33, 82, 83, 90, 92, 314
Bradley, P.S., 217, 315
Breslow, N.E., 192, 224, 315
Bridle, J.S., 315
Brockwell, P.J., 36, 75, 315
Broniatowski, M., 227, 315
Brookmeyer, R., 294, 315
Brooks, S.P., 238, 286, 315
Brownlee, K.A., 61, 62, 315
Buck, S.F., 30, 33, 45, 46, 315
Byrne, C.L., 32, 315
Byrne, W., 309, 315
- Cabrera, J., 310, 312
Cadez, I.V., 66, 315
Campillo, F., 284, 315
Carey, V., 310, 320
Carlin, B.P., 242, 315, 319
Carlin, J.B., 65, 123, 294, 313, 320
Carroll, R.J., 61, 332
Carson, R., 85, 325
Carter, W.H., Jr., 30, 315
Casella, G., 3, 34, 222, 223, 225, 228, 234,
 235, 237, 241–243, 249, 254–256,
 315, 325, 326, 332
Celeux, G., 227, 228, 284, 292, 315, 316, 332
Cepellini, R.M., 31, 33, 54, 316
Cerný, V., 284, 316
Chan, K.S., 222, 316
Chan, P.S., 173, 330
Chang, W.C., 33, 334
Chao, E.C., 276, 336
Chauveau, D., 228, 316
Cheeseman, P., 295, 316
Chen, J., 224, 316
Chen, K., 304, 305, 316
Chen, M.-H., 173, 323
Chen, T.T., 30, 33, 316
Cheng, B., 32, 295, 316

- Chernick, M.R., 130, 316
 Chi, H., 304, 316
 Chib, S., 223, 228, 240, 242, 245, 250, 311, 315, 316
 Christensen, R., 174, 316
 Clayton, D.G., 173, 192, 224, 238, 242, 315, 317, 320
 Coberly, W.A., 33, 331
 Cochran, W.G., 47–49, 317
 Cohen, M., 181, 317
 Congdon, P., 267, 317
 Constable, P.D.L., 88, 200, 312
 Cortijo, F., 57, 329
 Cowell, R.G., 36, 317
 Cox, D.R., 3, 317
 Cox, G., 47–49, 317
 Cramér, H., 3, 317
 Csiszár, I., 32, 309, 317
 Cuzick, J., 173, 317
 Dalal, S.R., 181, 317
 Datta, S., 275, 310, 314
 Daube-Witherspoon, M.E., 287, 317
 Davenport, J.W., 284, 286, 317
 Davidian, M., 224, 316
 Davidon, W.C., 152, 317
 Davis, R.A., 36, 75, 315
 Dawid, A.P., 36, 317
 Day, N.E., 17, 31, 64, 317
 De Gruttola, V., 331
 De Pierro, A.R., 287, 317
 DebRoy, S., 188, 191, 313
 DeGruttola, V., 173, 294, 317
 Delampady, M., xx, 233, 312
 Dellaert, F., 281, 317
 Demidenko, E., 191, 317
 Dempster, A.P., xix, xxi, xxii, 1, 2, 8, 13, 17, 19, 20, 24, 29–34, 66, 78, 90, 92, 100, 101, 121, 161, 182, 193, 216, 289, 291, 317, 318
 Dennis, J.E., 3, 317
 Didelez, V., 36, 318
 Diebolt, J., 227, 228, 250, 284, 292, 315, 316, 318, 332
 Dietz, E., 142, 314
 Dinse, G.E., 169, 325
 Do, K.-A., 14, 310, 318, 328, 335
 Doucet, A., 254, 312
 Doursat, R., 295, 320
 Duan, J.-C., 196, 318
 Duda, R.O., 33, 309, 318
 Dudoit, S., 310, 320
 Dutta, S., 333
 Eagon, J.A., 30, 291, 313
 Edwards, J.W., 310, 311
 Efron, B., 4, 30, 31, 79, 130, 131, 318
 Eggmont, P.P.B., 263, 318
 Elandt-Johnson, R., 51, 318
 Elashoff, M., 273–275, 318
 Elkan, C., 310, 313
 Enders, C.K., 267, 318
 Engle, R.F., 36, 75, 336
 Escobar, M., 250, 318
 Everitt, B.S., 3, 146, 286, 290, 318, 319
 Ewens, W.J., 310, 319
 Fairclough, D.L., 284, 319
 Farewell, V.T., 169, 319
 Farmer, J., 295, 327
 Fayyad, U.M., 217, 315
 Fessler, J.A., 57, 160, 203, 319
 Fienberg, S.E., 30, 102, 162, 314, 316, 319
 Fine, J.P., 310, 314
 Fisher, R.A., 8, 34, 79, 319
 Flury, B., 93, 94, 319
 Fokoué, E., 207, 319
 Ford, I., 54, 327
 Foulley, J.L., 188, 212, 319
 Foxall, R., 295, 311
 Freeman, D., 295, 316
 Friedman, J.H., 309, 322
 Frühwirth-Schnatter, S., 290, 319
 Galbraith, A.J., 172, 173, 328
 Gamerman, D., 238, 319
 Ganeshalingam, S., 32, 319
 Garrett, E.S., 310, 331
 Gelatt, C.D., Jr., 284, 324
 Gelfand, A.E., 239, 242, 319
 Gelman, A., 59, 65, 219, 231, 237, 246, 250, 320
 Geman, D., 242, 285, 293, 320
 Geman, S., 213, 242, 285, 293, 295, 320
 Geng, Z., 36, 320
 Gentle, J.E., 233, 324
 Gentleman, R., 310, 320
 George, E.I., 241, 242, 315
 Geweke, J., 234, 320
 Geyer, C.J., 237, 242, 320
 Ghahramani, Z., 276, 277, 313
 Ghosh, J., 336
 Gianola, D., 258, 334
 Gilks, W.R., 234, 238, 240, 242, 320
 Gill, R.D., 31, 173, 320, 330
 Godambe, V.P., 271, 273, 320
 Gold, M.S., 321

- Golden, R.M., 295, 321
 Golub, G.H., 144, 321
 Goodman, L.A., 33, 321
 Goodnight, J.H., 180, 321
 Gordon, N.H., 171, 321
 Graham, J.W., 267, 321
 Grant, G.R., 310, 319
 Green, M., 168, 328
 Green, P.J., 213–215, 238, 242, 294, 314, 321
 Greenberg, E., 240, 316
 Grenander, U., 242, 321
 Guihenneuc-Joyal, C., 238, 329
- Haberman, S.J., 30, 33, 82, 102, 321
 Hahn, F.J., 31, 333
 Hall, B.H., 115, 314
 Hall, R.E., 115, 314
 Hamilton, J.D., 73, 321
 Hand, D.J., 290, 309, 319, 321, 336
 Hart, P.E., 33, 309, 318
 Hartley, H.O., 8, 30, 33, 83, 102, 132, 321,
 322
 Harville, D.A., 182, 322
 Hasselblad, V., 17, 33, 322
 Hastie, T., 309, 322
 Hastings, W.K., 322
 Hathaway, R.J., 65, 84, 207, 284, 317, 322
 Hausman, J.A., 115, 314
 Healy, M.J.R., 30, 33, 47, 322
 Heckman, J., 146, 322
 Heiser, W.J., 278, 322
 Heitjan, D.F., 65, 66, 265, 267, 322
 Hennig, C., 207, 322
 Hero, A.O., 57, 160, 203, 319
 Herzberg, A.M., 168, 312
 Hesterberg, T., 234, 322
 Heyde, C.C., 35, 270–272, 320, 322
 Higdon, D., 238, 314
 Hills, S.E., 242, 319
 Hinkley, D.V., 3, 4, 317, 318
 Hinton, G.E., 216, 269, 281, 295, 306, 308,
 309, 311, 323, 329, 330, 337
 Hobert, J.P., 192, 222–225, 249, 314
 Hocking, R.R., 33, 83, 132, 188, 322
 Hofer, S.M., 267, 321
 Holland, P.W., 162, 314
 Holmes, A.P., 54, 327
 Holst, U., 290, 322
 Horng, S.C., 34, 102, 196, 323
 Hosmer, D.W., Jr., 33, 323
 Huang, Y., 276, 336
 Huber, P.J., 61, 62, 323
 Huber, W., 310, 320
- Hughes, J.P., 293, 323, 327
 Hunter, D.R., 20, 94, 99, 269, 278, 281, 323,
 325
- Ibrahim, J.G., 173, 323
 Ichimura, M., 36, 320
 Ikeda, S., 157, 323
 Ingrassia, S., 207, 285, 323
 Ip, E.H.S., 228, 318
 Irizarry, R.A., 310, 320, 331
 Iusem, A.N., 323
- Jacobs, R.A., 295, 302, 303, 307, 323, 324
 Jaffrezic, F., 188, 319
 Jamshidian, M., 34, 35, 122, 125, 131, 138,
 144–146, 323
 Jeffcoat, M.K., 276, 336
 Jennrich, R.I., 35, 122, 125, 131, 138, 144–
 146, 185, 323
 Jewell, N.P., 34, 214, 215, 294, 313, 333
 Ji, C., 295, 327
 Jiang, Q., 208, 337
 Jolliffe, I.T., 197, 323
 Jones, M.C., 213, 333
 Jones, P.N., 65, 67, 72, 116, 142, 143, 146,
 324, 328
 Jordan, M.I., 284, 295, 300, 302, 303, 306,
 307, 323, 324, 333, 337
 Jöreskog, K.G., 195, 324
 Jørgensen, B., 25, 286, 287, 324
 Juang, B.H., 291, 324
- Kale, B.K., 273, 320
 Karim, M.R., 242
 Kaufman, L., 56, 335
 Kay, J., 54, 213, 324
 Keith, J.M., 310, 324
 Kelly, J., 295, 316
 Kempthorne, O., 51, 324
 Kennedy, W.J., Jr., 233, 283, 324, 336
 Kent, J.T., 88, 200, 312, 324
 Khan, R.N., 293, 324
 Kiefer, N.M., 65, 75, 324
 Kim, D.K., 85, 136, 324
 Kim, J.-A., 173, 313
 King, G., 250, 320
 Kingman, J.F.C., 154, 324
 Kirby, A.J., 238, 242, 320
 Kirkpatrick, S., 284, 324
 Klein, J.P., 173, 324
 Knowlan, S.J., 323
 Kong, A., 243, 261, 327
 Korn, E.L., 310, 333
 Kotz, S., 58, 324

- Krishnan, T., xx, 57, 69, 233, 311, 312, 324, 325, 330
- Kumar, V., 336
- Kurata, K., 309, 312
- Kuroda, M., 36, 320
- Laarhoven, P.J.M., 335
- Lahiri, D.B., 233, 325
- Lai, T.L., 300, 325
- Laird, N.M., xix, xxi, xxiii, 1, 2, 8, 13, 17, 19, 20, 24, 29–34, 78, 90, 92, 100, 101, 121, 161, 182, 185, 216, 289, 291, 317, 318, 325
- Langari, R., 296, 297, 300, 325, 336
- Lange, K., 3, 6, 25, 35, 59, 61, 62, 85, 105, 149–153, 155–157, 176, 177, 214, 220, 231, 235, 242, 256, 269, 278, 279, 323, 325, 334
- Lange, N., 182, 325
- Lansky, D., 34, 150, 325
- LaRiccia, V.N., 263, 318
- Larson, M.G., 169, 325
- Lauritzen, S.L., 36, 317
- Lavielle, M., 228, 325
- Lawley, D.N., 193, 195, 196, 325
- Le Gland, F., 284, 315
- Ledolter, J., 222, 316
- Lee, A.H., 174, 192, 304, 326, 330, 337
- Lee, D., 217, 218, 287, 335
- Lee, M.-L.T., 310, 326
- Lee, T.-M., 242, 319
- Lehmann, E.L., 3, 326
- Leroux, B.G., 292, 326
- Levin, E., 295, 326
- Levine, R.A., 222, 249, 326
- Liang, K.-Y., 326
- Lindgren, G., 290, 322
- Lindsay, B.G., 6, 65, 142, 286, 290, 314, 326
- Lindstrom, M.J., 142, 326
- Little, R.J.A., xxii, 31, 41, 45, 51, 59, 125, 176, 177, 180, 197, 212, 265, 266, 289, 290, 313, 325, 326
- Liu, B., 336
- Liu, C., xxii, 35, 36, 59, 61, 159, 160, 175–178, 181, 182, 185–187, 196, 197, 201, 204, 212, 326, 327
- Liu, J.S., 235, 240, 243, 261, 262, 327
- Lopes, H.F., 238, 319
- Louis, T.A., 34, 35, 106, 111, 132, 137, 138, 142, 146, 225, 226, 327
- Loukas, S., 66, 311
- Lucy, L.B., 30, 57, 327
- Luginbuhl, T.E., 300, 334
- Lystig, T.C., 293, 327
- Ma, S., 295, 298, 304, 327
- MacEachern, S.N., 237, 327
- Madow, W.G., 290, 327
- Makov, U.E., 32, 335
- Mallows, C.L., 32, 312
- Mannila, H., 309, 321
- Mari, S., 309
- Marin, J.-M., 238, 327
- Mateos, J., 57, 329
- Matthews, D.E., 25, 149, 150, 332
- MaWhinney, S., 294, 331
- Maxwell, A.E., 193, 195, 196, 325
- McClure, D.E., 213, 320
- McClure, J., 310, 336
- McColl, J.H., 54, 327
- McCulloch, C.E., 34, 124, 223, 225, 247, 249, 325, 327
- McGiffin, D.C., 169, 172, 173, 328
- McGilchrist, C.A., 192, 327
- McKendrick, A.G., 29, 327
- McLachlan, G.J., xx, 4, 13, 14, 17, 31, 32, 63, 65–67, 72, 115, 116, 142–144, 146, 167–169, 172–174, 191, 192, 195, 204, 205, 207–209, 217, 289, 290, 292, 293, 296, 301, 304–307, 309, 310, 315, 318, 319, 324, 326–328, 330, 331, 336, 337
- McLaren, C.E., 66, 315
- McNeil, A.J., 238, 242, 320
- McShane, L.M., 310, 333
- Meilijson, I., 7, 34, 115, 120, 123, 132, 138, 139, 141, 328
- Meng, X.-L., xx, xxii, 29, 34, 35, 57, 100, 101, 108, 120–123, 125–131, 159, 160, 162, 163, 165, 166, 175, 181, 182, 198–204, 214, 224, 258, 263, 293, 294, 328, 329, 335
- Mengersen, K., 238, 314, 329
- Metropolis, N., 239, 329
- Minka, T., 281, 329
- Mitchell, T.J., 34, 329, 335
- Moeschberger, M.L., 173, 324
- Molenberghs, G., 191, 336
- Molina, R., 57, 329
- Mollie, A., 242, 314
- Monahan, J.F., 52, 329
- Moore, A.W., 217, 329
- Morgan, B.J.T., 286, 315
- Morton, R., 35, 270, 272, 322
- Mosimann, J.E., 156, 157, 329
- Motoda, H., 336

- Moulines, E., 228, 325
Muehllehner, G., 287, 317
Müller, P., 310, 335
Murray, G.D., 85, 181, 204, 329
Myers, R.H., 30, 315
- Nadarajah, S., 58, 324
Nagaoka, H., 309, 312
Narayanan, A., 156, 329
Nash, S.G., 144, 321
Neal, R., 269
Neal, R.N., 216, 281, 309, 329
Nelder, J.A., 102, 330
Nettleton, D., 85, 330
Newcomb, S., 330
Ng, H.K.T., 173, 330
Ng, S.K., xx, 172, 173, 191, 192, 217, 296,
 304–307, 310, 328, 330, 336, 337
Ng, S.W., 191, 330
Nielsen, G.G., 173, 330
Nowlan, S.J., 295, 330
Núñez, J., 57, 329
Nychka, D.W., 213, 331, 333
- Oakes, D., 35, 133, 134, 331
Olkin, I., 290, 327
Orchard, T., 30, 33, 34, 96, 331
Ord, J.K., 3, 334
Ostrowski, A.M., 83, 331
- Pagano, M., 294, 331, 335
Page, G.P., 310, 311
Parmigiani, G., 310, 331
Pawitan, Y., 187, 191, 331
Pedlow, S., xxii, 329
Peel, D., 13, 63, 65, 195, 204, 205, 207, 209,
 289, 290, 293, 301, 310, 328, 331
Peng, F., 295, 323
Peters, B.C., 33, 65, 331
Petrie, T., 30, 256, 291, 313
Pettitt, A.N., 32, 331
Peugh, J.L., 267, 318
Phillips, R.F., 32, 331
Pierce, M.A., 284, 317
Pierni, W.C., 284, 319
Pigeot, I., 36, 318
Pincus, M., 284, 331
Pinheiro, J.C., 188, 191, 313, 331
Polson, N.G., 242, 332
Powell, M.J.D., 153, 331
Prado, P., 168, 328
Propp, J.G., 238, 331
Puterman, M.L., 292, 326
- Qian, W., 32, 291, 293, 331
Quinlan, J.R., 336
- Rabiner, L.R., 32, 291, 292, 324, 331
Racine-Poon, A., 242, 319
Radmacher, M.D., 310, 333
Rai, S.N., 25, 149, 150, 332
Rao, C.R., 3, 8, 51, 332
Ratschek, H., 3, 332
Redner, R.A., 6, 7, 34, 115, 146, 290, 332
Reina, C.A., 217, 315
Richardson, S., 238, 320
Richardson, W.H., 30, 57, 332
Ridgway, G.J., 284, 319
Ripley, B.D., 238, 284, 285, 295, 310, 332
Ritter, C., 242, 332
Robert, C.P., 223, 225, 228, 234, 235, 237,
 238, 243, 250, 254–256, 292, 293,
 312, 318, 327, 329, 332
Robert-Granie, C., 188, 319
Roberts, G.O., 238, 242, 250, 256–258, 315,
 332, 333
Rokne, J., 3, 332
Rosenbluth, A.W., 239, 329
Rosenbluth, M.N., 239, 329
Roweis, S., 196, 197, 332
Rubin, D.B., xix, xxi–xxiii, 1, 2, 8, 13, 17, 19,
 20, 24, 29–36, 41, 45, 51, 59, 61,
 65, 66, 78, 90, 92, 100, 101, 108,
 120–123, 125–131, 159–162, 165,
 175–178, 180–182, 185–187, 193,
 195–197, 203, 204, 212, 214, 216,
 219, 230, 237, 246, 258, 265, 266,
 289–291, 313, 317, 320, 322, 326,
 327, 329, 332
Ruppert, D., 61, 332
Ruud, P.A., 36, 75, 332
Ryan, L., 273–275, 318
- Sahu, S.K., 257, 258, 333
Sarkar, A., 265, 333
Satten, G.A., 275, 333
Saul, L.K., 300, 333
Schader, M., 66, 333
Schafer, J., xxii, 333
Schaub, R., 142, 314
Schlattmann, P., 142, 314
Schlossmacher, E.J., 32, 333
Schluchter, M.D., 185, 323
Schmee, J., 31, 333
Schmid, F., 66, 333
Schnabel, R.B., 3, 317
Schneider, T., 197, 333

- Schwertman, N.C., 284, 319
 Scott, J.A., 192, 326
 Searle, R., 247, 327
 Segal, M.R., 34, 124, 214–216, 294, 310, 313,
 314, 333
 Sejnowski, T.J., 308, 311
 Self, M., 295, 316
 Sexton, J., 163, 164, 258, 333
 Sharples, L.D., 238, 242, 320
 Shepp, L.A., 56, 333, 335
 Shih, W.J., 204, 333
 Shumway, R.H., 36, 75, 333
 Silverman, B.W., 213, 294, 333
 Simon, R.M., 310, 333
 Simonato, J.-G., 196, 318
 Singer, B., 146, 322
 Sinha, D., 173, 323
 Siniscalco, S., 31, 316
 Sinsheimer, S.S., 59, 325
 Smith, A.F.M., 32, 238, 239, 242, 250, 256,
 319, 333, 335
 Smith, C.A.B., 31, 121, 316, 333
 Smyth, P., 66, 309, 315, 321
 Sobel, E., 242, 334
 Solla, S.A., 295, 326
 Sørensen, T.I.A., 173, 330
 Sørenson, D., 258, 334
 Soules, G., 30, 313
 Specht, D.F., 295, 334
 Speed, T.P., 310, 334
 Spiegelhalter, D.J., 36, 238, 242, 317, 320
 Steele, B.M., 224, 334
 Steinbach, M., 336
 Steinberg, D., 336
 Stern, H.S., 65, 320
 Stoffer, D.S., 36, 75, 333
 Stolorz, P., 284, 337
 Stork, G.E., 309, 318
 Stram, D., 182, 325
 Streit, R.L., 300, 334
 Stuart, A., 3, 334
 Stutz, J., 295, 316
 Sundberg, R., 31, 101, 102, 334
 Svensén, M., 207, 334
 Swensen, A.R., 163, 164, 258, 333
 Tan, W.Y., 33, 334
 Tanaka, J.S., 196, 313
 Tananbaum, H., 32, 312
 Tang, R., 212, 335
 Tanner, M.A., xxii, 132, 221, 222, 228, 230,
 242, 258, 295, 323, 332, 334, 336
 Tao, F., 36, 320
 Taylor, J.M.G., 59, 85, 136, 176, 177, 324,
 325
 Taylor, W., 295, 316
 Teller, A.H., 239, 329
 Teller, E., 239, 329
 Thayer, D.T., 193, 195, 196, 332
 Thiesson, B., 36, 334
 Thisted, R.A., 3, 8–10, 13, 14, 164, 334
 Thompson, E.A., 33, 102, 242, 320, 334
 Tibshirani, R., 130, 131, 309, 318, 322
 Tierney, L., 238, 334
 Tipping, M.E., 196, 197, 207, 335
 Tishby, N., 295, 326
 Titterington, D.M., 32, 63, 150, 207, 249, 277,
 287, 290, 291, 293, 295, 308, 309,
 312, 316, 319, 331, 335, 336
 Tsaiatis, A.A., 173, 336
 Tsitsiklis, J., 285, 314
 Tu, X.M., 173, 294, 317, 331, 335
 Tukey, J.W., 181, 317
 Turnbull, B.W., 30, 31, 34, 294, 329, 335
 Tusnády, G., 32, 309, 317
 Tyler, D.E., 200, 324
 Ulam, S., 239, 329
 Utans, J., 284, 337
 Vaida, F., 224, 335
 van Dyk, D.A., 29, 35, 57, 160, 175, 181, 188,
 198–204, 212, 263, 293, 319, 329,
 335
 van Laarhoven, P.J.M., 285, 311
 Vannucci, M., 310, 335
 Vardi, Y., 56, 85, 200, 217, 218, 287, 324,
 333, 335
 Vecchi, M.P., 284, 324
 Verbeke, G., 191, 336
 Walker, H.F., 6, 7, 34, 65, 115, 146, 290, 331,
 332
 Wan, K., 36, 320
 Wang, B., 277, 336
 Wang, C.Y., 276, 336
 Wang, K., 191, 192, 330
 Wang, L., 296, 297, 300, 325, 336
 Ware, J.H., 182, 185, 325
 Watson, I.F., 294, 313
 Watson, M., 36, 75, 336
 Wei, G.C.G., 132, 221, 222, 228, 230, 336
 Weisberg, S., 204, 333
 Weiss, G.H., 285, 336
 Weiss, N., 30, 313
 West, M., 250, 318
 Westmacott, M., 30, 33, 47, 322

- Wild, P., 234, 320
Wilson, D.B., 238, 331
Wilson, J.D., 213, 333
Wit, E., 310, 336
Wolfe, J.H., 17, 336
Wolynetz, M.S., 31, 336
Wong, S.P.S., 300, 325
Wong, W.H., 228, 243, 258, 261, 327, 334
Woodbury, M.A., 30, 33, 34, 96, 331
Wright, G.W., 310, 333
Wright, K., 283, 336
Wu, C.F.J., 25, 33, 80–84, 161, 336
Wu, X., 336
Wu, Y.N., 36, 327
Wulfsohn, M.S., 173, 336
- Xiang, L., 192, 337
Xu, L., 284, 295, 303, 304, 306, 307, 316,
 324, 337
Xu, R., 335
- Yang, I., 278, 325
Yang, Q., 336
Yates, F.Y., 144, 337
Yau, K.K.W., 174, 192, 326, 327, 330, 337
Yen, J., 296, 300, 325
York, J., 242, 314
Yu, P.S., 336
Yuille, A.L., 284, 337
- Zangwill, W.I., 164, 337
Zeger, S.L., 242, 310, 326, 331
Zhang, D., 224, 316
Zhao, J., 208, 337
Zhao, Y., 310, 333
Zhou, Z.-H., 336
Zoppé, A., 93, 94, 319

This Page Intentionally Left Blank

SUBJECT INDEX

- AECM algorithm, xxiv, 35–36, 160, 202–205, 209, 211–212, 263
- AEM algorithm, 145–146
- Accelerated EM algorithm, *see* AEM algorithm
- Acceleration of the EM algorithm, 35–36, 105, 136–145, 152–153, 156
- methods of
- Aitken, xxiv, 35, 137–138, 144
- conjugate gradient, xxiii, 35, 144–146
- Louis, 35, 137–138, 141–143
- quasi-Newton, xxiv, 25, 35–36, 138, 146, 150, 151–153, 156
- multinomial distribution example, 138–139
- geometric mixture example, 139–142
- grouped and truncated data, for, 142–143
- Aitken’s acceleration method, *see* Acceleration of the EM algorithm, methods of, Aitken
- Alternating expectation-conditional maximization algorithm, *see* AECM algorithm
- AR(1) model, *see* Hidden Markov model
- Baker’s method, *see* Covariance matrix of MLE
- Baum-Welch algorithm, 291
- Bayesian EM, 36–37, 230–232
- log posterior function, 231
- posterior mode by EM, 230–231
- Bayesian inference, xxiii, 3, 26–27, 108, 132, 173, 188, 191, 207, 219–220, 228–232, 238–242, 245–246, 249–254, 256, 258, 260, 265, 267, 292
- Gibbs sampler, use of, in, 241–245
- hidden Markov model, in, 292–293
- Best linear unbiased predictor (BLUP), 187, 190
- Beta distribution, *see* Distributions, beta distribution
- Bimodal data
- two-component normal mixture for, 65–66
- Binomial distribution, *see* Distributions, binomial distribution
- Bioinformatics, 310
- Bivariate normal distribution, *see* Distributions, normal, bivariate
- Boltzmann
- distribution, *see* Distributions, Boltzmann machine, *see* Neural networks
- Bootstrap
- covariance matrix estimation, for, 130–131. *See also* Standard error estimation, bootstrap approach to missing data problems, in, 130–131
- Buck’s method, *see* Examples, multivariate normal with missing values, Buck’s method
- Cauchy distribution, *see* Distributions, Cauchy distribution
- Censoring, 2, 33, 98–99, 169, 293
- Census undercount, xxiv, 2
- Chi-squared distribution, *see* Distributions, chi-squared distribution
- CM algorithm
- ECM algorithm in complete problem, as, 162
- CM-step, 35, 160–167, 172–175, 177–178, 185, 196, 202–204, 206, 261, 304
- definition of, 160–161
- Coarsened data, 265
- Compactness
- parameter space, of, 81, 84
- Comparison of algorithms, 284–286
- Complete-data
- information matrix, 96–99, 108–109, 114, 122, 125, 132
- likelihood, 2, 18, 38, 53, 60, 69, 115, 200, 209, 277
- log likelihood, 11–12, 15, 18–19, 21, 34–35, 38, 43–44, 50, 53, 56, 60, 63, 68–69, 74, 86, 94–97, 106, 108, 115, 118, 134, 155, 160, 170, 175, 177, 183, 192, 194, 199, 203, 206, 209, 212, 221, 223, 225–227, 231, 248, 252, 259, 263, 272, 277, 283, 295–297, 299, 301–302
- conditional expectation of, 11, 16, 19, 21, 23, 26, 28, 36–37, 43, 50, 60, 74, 86, 93–95, 134, 160, 170, 176–177, 183, 200, 202–203, 206, 221, 231, 277, 283, 295–296, 299
- gradient, 34, 108, 112, 115
- curvature, 34, 108
- problem, 2, 28, 33, 35, 50, 106, 164, 197, 271
- specification

- choice of, 2, 20, 33, 35, 50, 77, 106, 108, 198, 271, 276
 factor analysis, in, 2
 score statistic, 34, 38, 95, 114
 Dirichlet distribution example, 156
 single observation, 116, 140
 summary statistic 273
 vector, 11, 15, 18, 20–21, 27, 37, 43, 50, 52, 56, 59, 63, 68, 118, 139, 154, 182, 192–193, 199–200, 202–204, 223, 225, 248, 250, 252, 259, 273
- Conditional maximization algorithm, *see CM algorithm*
- Conjugate gradient algorithm, 144–145
- Conjugate prior, 60, 251
- Constrained parameter spaces, 84–85
- Contingency tables, 30, 289–290
 incomplete data, with, 32, 102
 ECM algorithm for, 174–175
- Convergence
 AECM algorithm, of, 203
 EM algorithm, of, xix, xxiii, 19–20, 28–29, 33–34, 77, 79–90, 143, 161–162, 177
 grouped data, 120
 EM gradient algorithm, 149
 GEM algorithm, of, 81–84
 ECM algorithm, of, 162–165
 ECME algorithm, of, 175
 diagnostics for MCMC methods, 237–238
 EM in factor analysis, of, 102, 195–196
 faster
 ECME over EM, 181–182
 geometric rate, Gibbs sampler, of, 242
 Gibbs sampler, of, 257–258
 linear, 13, 22, 33, 100–101, 124, 143–144
 local linear, 7
 OSL algorithm, of, 214
 quadratic, 5–6, 9
 rate, *see Rate of convergence*
 slowness of, xix, xxiii, 2, 29, 33, 99–102, 105, 120, 124, 142, 177, 195–196
 ECM relative to EM, 159, 163, 165
 ECME relative to ECM, 175–176
 multicycle ECM relative to ECM, 166
 SA relative to EM, 285
 speed, 20, 35, 99–100, 106, 150, 159, 198–199, 201, 212
 ECM algorithm, of, 162–163, 165, 175–175
- ECME algorithm, of, 176, 181
 sublinear, 34, 102, 196
 factor analysis, in, 102
 superlinear, 6, 164–165
- Convolutions
 EM algorithm in, 2
- Covariance components in longitudinal data analysis, 188
- Covariance matrix of MLE, xxiii, 2, 4, 34, 52, 73, 105–106, 108, 116, 121, 132, 142, 173, 226. *See also Standard error estimation*
 asymptotic, 4, 30, 34, 120–121
 Supplemented EM algorithm, via, 121
- Baker's Method, 131–132, 135–136
 Standard error for categorical data, 135
 computation of, 34
 empirical, 34, 120, 141
 geometric mixture example, 146
 estimation of, 4, 6, 29, 105, 108, 116, 143
 grouped and truncated data, for, 73
- Expectation-Solution Algorithm, for, 275
 inverse of, 4
- Cox Model, 173
- Cyclic coordinate ascent method
 CM algorithm as special case of, 164
- Dairy science, xxii, 2
- Data augmentation algorithm, 228–230
 chained, 230
 efficient, 198–202
 multivariate *t*-distribution example, 198–202
 rate of convergence of, 198
 variance components model, for, 202
 working parameter, 198–202
- EM and MI, as a combination of, 265
- Gibbs sampler as extension of, 262
- imputation step (I-step), 229
- Poor man's (PMDA)
 PMDA-1, 229–230
 PMDA-2, 230
 posterior step (P-step), 229
- Data mining, 309–310
- Data sets
 ABO blood group data, 51
 number of lettuce plants from a 3² experiment, 47–48
 bivariate data with missing values, 45, 85, 129
 Fisher's *Iris* data, 168

- genetic linkage data, 8
 red blood cell volume data, 65–67, 72–73, 142–143
 data on sampler's errors on shoot heights, 49
 serum protein data of Pekin ducklings, 156–157
 stack loss data, 61
 winter wheat data, 185–186
- Deconvolution problem, 30, 57, 85
 Delta algorithm, 284, 286–287
 Designed experiments, *see* Examples, least squares with missing data
 Digamma function, 155, 176, 211
 Dirichlet distribution, *see* Distributions, Dirichlet distribution
- Distributions**
- beta distribution, 27, 230
 - binomial distribution, 11–12, 53, 134, 226
 - Boltzmann distribution, 284
 - Cauchy distribution, 201
 - chi-squared distribution
 - testing regression errors, for, 75
 - testing goodness of fit, for, 61
 - Dirichlet distribution, 153–156
 - prior, use as, 250, 260
 - exponential distribution, 20–22, 117–120, 201
 - gamma distribution, 58–60, 153, 155, 176, 198, 208–210
 - geometric distribution, 139–142
 - Gibbs distribution, simulated annealing, in, 285
 - inverted gamma distribution, 259
 - log normal distribution, 73, 142
 - multinomial distribution, xxiii, 8–13, 36, 51–54, 68, 116, 133, 136, 138, 149, 215–216, 224, 226–229, 260, 274, 280, 283, 297–298, 301, 303, 307
 - negative binomial distribution, 66, 69
 - normal distribution
 - bivariate, 42–44, 86, 90, 93, 128–130, 163, 173, 181, 243, 262
 - multivariate, 31, 33, 36, 39, 41, 46, 58, 64–65, 102, 166, 168, 178–179, 182–183, 191, 194, 197, 204, 207–208, 258, 262, 267, 295, 298, 310
 - univariate, 17, 29, 33, 39, 42, 61–66, 81, 84–85, 111, 125, 201, 257, 285
 - Poisson distribution, 55–56, 66
- t*-distribution, *see also* Multivariate *t*-distribution
 ECME algorithm for, 178, 181–182
 error distribution as, 32, 62
 MLE of location parameter, 58–62
 multivariate, 58–62, 176–182, 198–202, 207–210, 212
 univariate, 88–90, 181, 212
 uniform distribution, 27, 94, 221–222, 233, 235, 285, 305, 308
 Wishart distribution, 263
- Downweighting
 outliers, of, 59
- ECM algorithm, xxiv, 27, 35–36, 159–166, 168, 172–178, 180–181, 202–205, 258
 convergence, 162, 212
 speed of, 162–165, 175, 181
 definition of, 160–161
 EM algorithm as special case of, 161
 Gibbs sampler analog, of, 261–262
 Hidden Markov model, for, 164
 mixture of experts, in, 304–305
 - comparison of ECM and IRLS, 305
 Multicycle ECM algorithm, *see* Multicycle ECM algorithm
 space-filling condition, 161–162
- ECME algorithm, xxiv, 35, 59, 159–160, 175–178, 180–182, 185–188, 190, 202–203,
 extension of ECME algorithm, 178
 factor analysis, in, 196
 Gibbs sampler analog, of, 263
 multivariate *t*-distribution, in, 59, 176, 178, 180–181
 variance component analysis, in, 185–186
- Efficient data augmentation, *see* Data augmentation algorithm, efficient
 e -geodesic, 32
 EM algorithm, *see also* E-step; M-step
 convergence of, xix, xxiii–xxiv, 19–20, 28–29, 33–34, 77, 79–85, 88–90, 143, 162, 177
 formulation of, 18–20
 Gibbs sampler analog, 262–263
 Healy-Westmacott procedure, as, 49–51
 history, 29–36
 MAP estimation, for, 26–27
 modifications, of, 57
 monotonicity, xxiii, 78–79, 80, 82
 MPL estimation, 27–28, 214–215

- nontypical behavior, 85–95
- rate of convergence of, *see* Rate of convergence
- self-consistency of, xxii, 30–31, 34, 77, 79, 93, 96
- speeding-up convergence of, 20, 25, 35, 142
- time series, in, 36, 73–75, 289–290, 300
- EM gradient algorithm, xxiv, 26, 36, 38, 149–152
 - Dirichlet distribution example, 156–157
- EM map, 26, 31, 82, 88, 90, 122, 215
 - fixed point, 36, 88, 90, 100
 - Jacobian matrix, 100
 - point-to-point, 81–82
 - point-to-set, 80–81
- EMS algorithm, 213
 - AIDS epidemiology, in, 294
 - penalized likelihood, as, 213
 - PET, for, 213
 - smoothing step in, 213
 - SPECT, for, 213
- Empirical Bayes and EM, 263–264
- E-step (Expectation step), 1–2
 - definition of, 19
 - MAP, for, 26–27
 - Monte Carlo, 29, 191, 224
 - multicycle E-step, 172
 - numerical integration in, 173
 - pathologies, 93–95
 - Q -function, 12, 24, 27, 38, 44, 139, 148, 159, 161–163, 165–166, 175–176, 182, 200–202, 225, 231, 249, 269, 282, 287, 296, 299, 302, 304
 - approximation by Monte Carlo E-step, 224
 - definition, 11
 - derivative, 162
 - ECM algorithm, in, 162
 - GEM algorithm, of, 24
- regular exponential family, for, 22–23
- stochastic, 227–228
 - multinomial, for, 227–228
- Estimating equations, 37, 269–271, 273–276
 - expected, 276
 - unbiased, 271
- Estimating function, 36, 270, 272, 275
 - optimal unbiased estimating function – conditional score function as, 270
 - optimal unbiased estimating function – score function as, 270
- Estimation
 - maximum likelihood, xxi–xxiii, 1, 3–4
 - maximum likelihood estimate, 2
 - nonparametric, 34
 - uniqueness of, 84
- maximum penalized likelihood, xxii, 26–28, 212, 214
 - AIDS epidemiology, in, 294
 - roughness penalty in, 28, 213
 - smoothing parameter in, 28
- maximum *a posteriori* (MAP), xxiii, 26–27, 37, 102, 212–213, 219, 231, 277, 283, 285
 - multinomial distribution example, 27
- mixing proportions, of, 13–16
- robust, xxii, 32, 59, 61, 270
- variance for MPL estimate, of, 214–216
- Examples
 - AR(1) model, 73–76, 164
 - Bayesian EM for normal with semi-conjugate prior, 231–232
 - Bayesian probit analysis with data augmentation, 250–251
 - bivariate normal distribution
 - missing data in, 42–46, 129–130
 - censored exponentially distributed survival times, 20–24, 97–99
 - contingency tables with incomplete data, 174–175
 - Data augmentation, 229–230
 - Data augmentation and Gibbs sampler for censored normal, 259–260
 - Dirichlet distribution, 153–156
 - EM-Gibbs connection for censored data from normal, 256
 - EM Gibbs connection for normal mixtures, 257
 - Expectation-Solution algorithm (ES)
 - multinomial, for, 274–275
 - exponential distribution
 - grouped data, from, 117–120
 - factor analysis, 192–197
 - genetics
 - genetic linkage estimation, 8–13
 - gene frequency estimation, 8, 51–54
 - geometric mixtures, 139–142
 - Gibbs sampling for bivariate normal, 243–245
 - Gibbs sampling for censored normal, 251–254
 - Gibbs sampler for a complex multinomial, 260–261
 - grouped and truncated data, 66–73
 - grouped log normal data, 72–73

- hidden Markov models, *see* Examples, AR(1) model
- Interval-EM algorithm
 - complex multinomial, for the, 283
- least squares with missing data
 - Latin square with missing data, 49
 - linear regression with missing data, 47–48
- log linear model
 - ECM algorithm for 174–175
- mixing proportions, estimation of, 13–18
- Monte Carlo EM
 - censored data from normal, for, 223
 - Oakes' standard error for a one-parameter multinomial, 226
 - two-parameter multinomial, for, 224
- Monte Carlo integration, 221
- M-H algorithm for Bayesian probit regression, 245–246
- MM algorithm
 - complex multinomial, for the, 280–281
- multinomial distribution, 8–13, 27–28, 51–54, 109–111, 138, 215–216, 224, 226, 259–261, 280–281, 283
- multivariate normal distribution with missing values
 - Buck's method, 45–46
- non applicability of EM algorithm
- normal mixtures
 - equal correlations, ECM example, 166
 - multivariate, 64–65
 - univariate, 17–18, 61–64, 111–113
- PET data, 54–58
- REML estimation in a hierarchical random effects model, 189–190
- repeated measures, *see* Examples, variance components
- SPECT data, 54–58
- survival analysis
 - ECM algorithm for mixture models in, 168–173
- t*-distribution
 - known degrees of freedom, with, 58–61
 - multivariate, 58–61, 176–182, 198–201
 - univariate, 88–92
 - unknown degrees of freedom, with, 176–182
- univariate contaminated normal data, 125–128
- variance components, 182–186
- Expectation-conditional maximization algorithm, *see* ECM algorithm
- Expectation-conditional maximization either algorithm, *see* ECME algorithm
- Expectation–maximization algorithm, *see* EM algorithm
- Expectation-Solution (ES) algorithm, 269, 273–275
 - multinomial, for, 274–275
- Expectation step, *see* E-step
- Exponential distribution, *see* Distributions, exponential
- Exponential family
 - curved, 82
- ECM algorithm for
 - convergence of, 162
- factor analysis model as a member of, 194
- hidden Markov model in, 292–293
- irregular
 - expected complete-data information matrix, 109
 - multinomial distribution example, 110
 - univariate contaminated normal model, 126
- linear
 - Poisson distribution, 56
- regular, xxiii, 21–23, 108–110, 122, 127–129, 148–150
 - bivariate normal distribution, 43
 - complete-data information matrix for, 122
 - Dirichlet distribution, 153–154
 - E- and M-steps for, 22–23, 53–54
 - EM gradient algorithm for, 150–151
 - expected information matrix for, 23
 - multinomial distribution, 53
 - natural or canonical parameter for, 22
 - observed information matrix for, 108
 - Q*-function for, 23
 - sufficient statistics, 22–23, 43, 53, 99
 - Sundberg formulas in, 31, 34
 - univariate normal distribution, 112
 - variance component model, in, 184
- type I censored data in, 30–31
- Extensions of EM algorithm, *see* ECM algorithm; ECME algorithm; EMS algorithm; MCEM algorithm; ACEM algorithm
- Factor analysis, xxiii, 37, 148, 193–197, 289
- AEM algorithm in, 148

- complete-data
 - problem for, 2
 - sufficient statistics for, 194
- EM algorithm in, 193–196
- ECME algorithm in, 196
- multiple solutions in, 195
- slow convergence of EM, 102, 195
- sublinear convergence of EM, 196
- Finite mixtures, 13, 32–33, 36, 72, 139–140, 168–169, 204, 227, 248, 281
 - contaminated model as special case, 126
 - geometric components
 - Aitken acceleration, 139–142
 - Gibbs sampler in, 249
 - Gompertz components, 171–172
 - hidden Markov model as extension of, 290–293
 - hierarchical mixtures, 295
 - hybrid methods, 146–147, 285–286
 - identifiability, 63
 - log normal components, 72–73
 - logistic regression-based mixing proportions, 169, 171
 - mixing proportions, estimation of, 13–18, 33, 63, 169
 - neural networks, in, 295, 301
 - normal components
 - comparison of algorithms, 285–286
 - comparison of EM algorithm and simulated annealing, 285
 - multivariate, 33, 64–65, 168
 - multivariate with equal correlation, 166
 - univariate, 17, 29, 33, 61–64, 72–73, 81, 84, 102, 111–113, 126
 - number of components, 144
 - Poisson components, 66
 - software for
 - EMMIX, 310
 - EMMIX-GENE, 310
 - EMMIX-WIRE, 191, 310
 - Stochastic EM algorithm in, 227–228
 - survival analysis, for
 - ECM algorithm, 168–174
 - Fisher's method of scoring, 3–10, 13, 28–29, 34, 62, 122, 156–157
 - Delta algorithm and, 286
 - Fisher's Z transformation, 129
 - Frailty model, gamma, 173
 - Gamma distribution, *see* Distributions, gamma distribution
 - Gauss-Newton algorithm, 8
 - Gauss-Seidel iteration method
 - ECM algorithm, as, 164
 - Generalized conjugate gradient algorithm, xxiv, 35, 144–145
 - Generalized EM algorithm, *see* GEM algorithm
 - GEM algorithm, xxiii, 24–26, 28, 83, 161
 - based on one NR-step, 24–25, 147–149
 - simulation experiment, 149
 - convergence of, 24, 81–82
 - monotonicity, of, 24, 79
 - nonconvergence, example of, 90–92
 - Generalized conjugate gradient algorithm, 144–145
 - Generalized linear mixed models, 191–192
 - MCEM, 224–225, 247
 - multivariate t importance sampling approximation, 191, 224–225
 - rejection sampling, 224–225
 - semiparametric approach, 224
 - two-slice EM, 224
 - Genetics, *see* Examples, genetics
 - gene expression microarray data, 310
 - motif patterns in DNA sequences, 310
 - Geometric distribution, *see* Distributions, geometric distribution
 - Gibbs distribution, *see* Distributions, Gibbs distribution
 - Gibbs sampler, 37, 220, 222, 241–263, 265, 293
 - analogs for EM, ECM, and ECME algorithms, 261–263
 - variance components estimation, for, 262
 - Bayesian computations, for, 242
 - case of full conditionals, 241
 - collapsing in, 261
 - complex multinomial, for a, 260–261
 - convergence diagnostics for, 237
 - finite mixtures, for, 249
 - geometric convergence of, 242
 - hidden Markov models, for, 293
 - homogeneous Markov chain, 244
 - marginal density, approximation of, 241, 244
 - Markov sequence, generating, 241
 - Metropolis-Hastings, special case of, 241
 - rate of convergence of, 257–258
 - variance components, for, 262
 - working parameter concept in, 263
 - Gibbs sequence, 241
 - Grouped and truncated data, 2, 66–73
 - incomplete-data score statistic, 70–71
 - LININPOS problem, as 217–218

- normal distribution, from, 71–72
- observed information matrix, 116–120
- Grouping**, 289
 - discrete data, in, 66
 - exponential data, in, 117–120
 - modified Newton-Raphson method for, 119–120
 - log normal data, in, 72–73
 - normal data, in, 71–72
- Hazard rate, semiparametric estimate of, 173
- Healy-Westmacott procedure**, 47–51
 - EM algorithm, as, 30, 49–51
- Henderson's algorithms**, *see* Variance components analysis, Henderson's algorithms
- Hessian (matrix)**, 5–8
- Hidden Markov models**, xxiv, 30, 37, 73–75, 164, 289–293
 - EM algorithm for, 30, 32, 290–293
- Hybrid algorithm**, xxiv, 146–147
 - Simulated annealing and quadratic programming, 285
- Ignorability**
 - missing data, of, 265–267
 - sufficiency based conditions for, 265
- I.I.D Monte Carlo algorithms**, 37, 220, 232–236
- Ikeda acceleration**, 157
- Image processing**, 30, 32, 54, 85, 217–218, 240, 242, 289–291, 293
 - MAP estimation in, 285
- Image space reconstruction algorithm (ISRA)**, 284, 287
 - PET, for, 287
- Importance sampling**, 220, 224–225, 230, 234–235, 241, 275
 - sampling importance resampling algorithm (SIR) 230, 235
- Importance ratio**, 235
- Imputation**, *see* Multiple imputation
 - imputation and deletion of missing values, 30, 94, 265–267, 271
 - methods, 265–267
- Imputation step (I-step)**, 229–230, 264
- Incomplete data**
 - contingency tables, in, 30, 102, 174–175
 - information matrix, 96, 108, 132
 - likelihood, 12, 16, 19, 30, 35, 45, 50, 57, 64, 68–70, 78, 96–97, 106, 113, 117, 132, 152, 159, 175, 182, 185, 195, 202, 255–256, 272, 287
- problem**, 1–2, 18, 26, 30–31, 33, 128, 219, 254, 265, 272, 284, 287, 289
- score statistic**, 34–35, 38, 70, 95, 116, 119, 138, 140, 271, single observation, 34
- grouped exponential data**, 117–120
- traditional missing data**, 18
- vector**, 11
- Incremental EM algorithm (IEM)**, 216–217, 304
- sparse incremental EM**, 217
- Information geometry**
 - EM algorithm and, 32, 295
- Information matrix**
 - complete-data information matrix, 96–99, 114, 125, 132
 - expectation of, 108–109
 - regular exponential family, for, 122
 - irregular exponential family, for, 109
 - empirical information matrix, 5, 7–8, 34, 73, 114, 116–117, 140–142, 146
 - approximation to expected information matrix, 5–7
 - approximation to observed information matrix, 7
 - definition of, 7
 - geometric mixtures, for, 140
 - grouped data, for, 142
 - method of scoring, in, 7
 - observed, 116
 - single observation, 114–116
 - expected Fisher information matrix, 4, 52, 96
 - definition of, 4
 - i.i.d. case, 114
 - regular exponential family, for, 23
 - rate matrix, and 101–102
- fraction of missing information**
 - rate of convergence, relation to, 101
- observed information matrix**, xxiii, 4, 6–7, 34–35, 38, 105–113, 121–125, 128, 130–133, 135–136, 214, 226, 293. *See also* Grouped and truncated data
- approximations to**, 114–117
- computation of**, 108
- geometric mixtures**, for, 141
- grouped and truncated data**, for, 116
- regular exponential family**, for, 108
- Supplemented EM**, via, 108, 123
- standard error estimation**, use in, 4, 6, 34, 106

- missing information matrix, 39, 96–99, 106
- Interval EM algorithm, 283–284
 - interval function, 283
- Iterative proportional fitting (IPF)
 - convergence, 162
 - ECM algorithm as generalization of, 174–175
 - neural networks, in, 309
- Iterative simulation techniques, xxii, xxiv, 59, 219–267
 - extensions of EM algorithm, as, xxii, 220–228, 246–249, 254–258, 261–263
 - MAP estimate, for, 219
 - summarize posterior distributions, to, 59, 220–221
- Iteratively reweighted least-squares (IRLS)
 - EM algorithm, as, 31–32, 61
 - Mixture of experts, in, 302–7
 - comparison of IRLS and ECM, 305–306
- Jacobian, *see also* EM map, 22, 100, 215
 - Gibbs sampler, of, 258
 - numerical differentiation, by, 122–123, 214
 - componentwise convergence rate, using, 122
- Jensen’s inequality, 79, 277, 279, 282
- Kalman filtering, 75
- Kullback–Leibler divergence, 257, 264, 300
- Latent structure analysis, 2, 102
 - latent class model, 2
 - latent variables, 2, 193, 204, 207, 209, 220, 232, 277, 293
- Latin squares, 49
- Least squares, *see also* Examples, least squares
 - with missing data
 - missing data, in, 47–51
 - nonlinear
 - Gauss–Newton methods for, 8
 - trimmed, 61
- Likelihood
 - complete-data, 2, 18, 38, 53, 60, 69, 115, 200, 209, 277
 - incomplete-data, 12, 19, 45, 57, 78, 152, 255–256
 - global maxima, 29, 33, 59, 77
 - local maxima, 3, 6, 29, 65, 80, 83, 85, 88, 102, 124, 147, 168, 195
 - multiple, 65, 88
- spurious, 168
- local minima, 5–6, 77, 88–89
- log likelihood
 - concave, 5
 - regular exponential family, of, 22–23
 - unimodal, 3, 33, 80, 84
- unbounded, 4, 65
 - finite mixtures, in, 4, 65
 - many spikes, with, 59
- penalized likelihood, xxiii, 27–28, 212–213, 293–294
 - saddle points of, 5–6, 77, 80, 85–88, 124
- Linear inverse problems
 - positive linear inverse problems, 217–218, 287
 - image space reconstruction algorithm for, 287
- Linear mixed model, xx, 37, 186–192, 310
 - REML estimation for, 188–190
- Linear regression, 32, 46–47, 61, 180, 251, 266, 270. *See also* Examples, linear regression with missing values; Missing data, linear regression, in missing dependent values, 47–51
- LININPOS, *see* Linear inverse problems
- Log likelihood, *see* Likelihood
- Log linear model, *see* Examples, log linear model
- Log normal distribution, *see* Distributions, log normal distribution
- Louis’ method, 34, 132–135, 137–138, 142, 173, 226
- Lowener ordering, 286
- Lower bound algorithm
 - Delta method, as special case of, 286
- Lower bound maximization, 269, 281–283
 - optimization of surrogate function, 281
- Mahalanobis distance, 58–59
- Majorize-Minimize (MM) algorithm, 37, 269, 278–281
 - geometry of, 94, 278
 - MM as an extension of EM algorithm, 94
- Majorizing function, 278–280
 - methods for constructing, 279–280
 - relation to GEM, 279
- MAP estimate, *see* Estimation, maximum *a posteriori* estimate
- Markov chain
 - Boltzmann machine generating, 308
 - Gibbs sampler sequences as, 242
 - hidden Markov chain, 30, 32, 290–293

- Gibbs sampler for, 293
- simulated annealing, in, 285
- Stochastic EM estimates as, 228
- Markov chain Monte Carlo methods, xix–xx, xxii, 36–37, 191, 220, 223–224, 236–264
 - dependence on starting values, 237
 - exact sampling or perfect sampling, 238
 - EM-type algorithms, relationship with, 254–258, 261–263
 - Monte Carlo error in, 221–222, 246, 249
- Markov processes
 - simulated annealing, in, 284–285
- Markov random field
 - hidden Markov models, in, 290, 292
- Maximization-step, *see* M-step
- Maximum *a posteriori* (MAP) estimation, *see*
 - Estimation, maximum *a posteriori* (MAP)
- Maximum likelihood estimation, *see* Estimation, maximum likelihood estimate
- Maximum penalized likelihood estimation, *see*
 - Estimation, maximum penalized likelihood
- MCEM algorithm, 221–228, 249, 276, 300
 - E-step, 219
 - generalized linear mixed models, in, 224–225
 - monitoring convergence, 222–224
 - monotonicity, 222
- MCMC or (MC)² methods, *see* Markov chain Monte Carlo methods
- M-estimates
 - regression, in, 61
- Medical imaging, xxii, 2, 289
- Metropolis algorithm, 239–241, 284
 - Markov field structure of, 239
- Metropolis-Hastings algorithm, 37, 239–242, 247, 249
 - acceptance probability, 239–241, 246
 - image processing, 240, 242
 - posterior distributions, 240–241
 - spatial statistics, use in, 242
- m*-geodesic, 32
- Minorize-Maximize(MM) algorithm, *see*
 - Majorize-Minimize algorithm
- Minorizing function *see* Majorizing function
- Missing data
 - bootstrap for standard error estimation, 130–131
 - choice of, *see* Complete-data, specification, choice of
 - contingency tables, in, 174–175
 - slow convergence of EM, 102
 - designed experiments, in, xxiii, 33, 47–51, 102
 - estimation, 30, 45
 - Latin squares, 49
 - estimation of, 42
 - Gibbs sampler, in, 249
 - ignorability of, 265–267
 - validity of frequentist and Bayesian inferences, 267
 - least squares, in, 47–51
 - linear regression, in, 47–48
 - mechanism, 265–267
 - missing at random (MAR), 42, 179, 181, 265
 - missing completely at random (MCAR), 85, 265
 - normal distribution, in
 - AEM algorithm for, 145
 - bivariate, 42–45
 - multivariate, 30–31, 41, 45–46, 146
 - specification, 35, 68, 106
 - t*-distribution, multivariate, in, 181
 - traditional sense, in, 1–2, 18, 41, 289
 - nonparametric bootstrap for, 131
 - vector, 18, 21, 43, 59, 68, 182, 198, 227, 249, 290
- Missing information principle, 30, 33–34, 36, 77, 95–99
- rate of convergence, relation to, 22, 33, 77, 101, 121, 198, 212
- Missing value, *see* Missing-data, vector
- Missing variable, *see* Missing-data, vector
- Mixing proportions
 - estimation of, 13–18
 - logistic regression-based model, in, 168–173
 - unknown, 13, 62–63
- Mixed models, posterior mode in
- Mixtures, *see* Finite mixtures
- Mixtures of factor analyzers, 204–211, 310
 - AECM for, 205, 209, 211
 - dimensionality reduction, 204
 - normal distribution of, 204–205
 - t*-distribution of, 207–210
- MLE, *see* Estimation, maximum likelihood estimate
- Mode
 - posterior distribution, of, *see* Posterior, mode
- Monotonicity, *see also* EM algorithm, monotonicity
- EM algorithm, of, xxiii, 78–80

- GEM algorithm, of, 24, 79
- MCEM algorithm, of, 222
- Monte Carlo EM, *see* MCEM algorithm
- Monte Carlo E-step, *see* E-step, Monte Carlo
- Monte Carlo methods, *see also* Markov chain
 - Monte Carlo methods
 - bootstrap, for, 130–131
 - integration, 37, 220–222, 235
 - optimization, 222
- M-step, 2
 - definition of, 19
 - ECM algorithm, for, 28
 - GEM algorithm, for, 28
 - Gompertz mixtures, for, 171–172
 - grouped normal data, for, 70
 - MAP, for, 26–27
 - regular exponential family, for, 23
- Multicycle ECM algorithm, 165–166, 172–173, 178, 203–204
- Multinomial distribution, *see* Distributions, multinomial distribution
- Multiple imputation (MI), 37, 230, 264–265, 267
- Multiresolution kd-trees, acceleration of EM, 217
- Multivariate normal distribution, *see* Distributions, normal, multivariate
- Multivariate normal mixtures, *see* Finite mixtures, normal components, multivariate
- Multivariate *t*-distribution, *see also* Distributions, *t*-distribution, multivariate
 - known degrees of freedom, 58–61
 - unknown degrees of freedom, 176–182
- Negative binomial distribution, *see* Distributions, negative binomial distribution
- Neural networks, xxiv, 32, 37, 289, 295–310
 - Boltzmann machine, 32, 295, 308–309
 - Iterative proportional fitting algorithm, 32, 309
 - probability model for, 295
- CM-step in, 304
- EM algorithm in, 37
- Hierarchical Mixture of experts, 307–309
 - multinomial distribution, 308
- hidden variables in, 296, 300, 308
- Mixture of experts, 296, 301–304
 - generalized Bernoulli density, 304
 - normalized mixture of experts, 306–307
- multiclass classification, for, 296, 301, 303–304
- Multi-layer perceptron networks, 297–300
- E-step, intractability of exact, 301
- normalized exponential function, 298
- radial basis function (RBF), 297, 301–302
 - clustering to find centres, 301
- training of, 295–303, 305–306
- Newton-type methods, 5–8, 28–29, 34, 105
- Newton-Raphson method, xxiii, 1, 3–9, 13, 24–25, 28, 35, 117, 119–120, 138, 141–142, 146–156, 167, 171, 173, 188, 190, 195–196, 296, 304–305
 - Delta algorithm as modification of, 286
- quasi-Newton methods, xxiv, 3, 5–7, 25, 35–36, 138, 146, 150–157, 164, 167, 172, 185, 285
- Dirichlet distribution example, 156–157
- modified Newton methods, 5–8
- variants, 5
- Non-applicability of EM algorithm, 41
- Nonconvergence
 - GEM sequence, of a, 81–82
- Nonlinear least squares, 8
- Nonparametric MLE, *see also* Estimation, maximum likelihood estimate, nonparametric
 - survival analysis, in, 30
- Normal distribution, *see* Distributions, normal distribution
- Normalizing constant, specification only up to, 232
- Numerical differentiation
 - Jacobian calculation by, 122–123
- Numerical integration, *see also* Supplemented EM
 - E-step, in, 173
- One-step-late algorithm, *see* OSL algorithm
- OSL algorithm, 28, 213–215
 - convergence of, 214
 - MAP, relation to, 213
 - MPLE, relation to, 28, 213
 - MPLE variance estimation, for, 215
 - PET, in, 213
 - smoothness parameter, in, 213
 - roughness penalty, in, 213
 - SPECT, in, 213

- Oakes' formula *see* Standard Error, 133
 example, 134
- Outliers, 59, 207
- Parameter expanded EM algorithm (PX-EM),
 201, 212
 data augmentation and ECME in mixed
 models, 188
 maximum likelihood and, 212
 working parameter, 212
- Penalized EM algorithm
 EMS algorithm, as, 213
 MPLE variance estimation, for, 214–215
- PET, 54–58, 85, 286, 289, 294
- AECM algorithm for, 204
 EMS algorithm for, 213
 detector, 55
 LININPOS problem, as, 85, 218
 OSL algorithm for, 213
 Sampling-importance resampling in, 230
- Poisson distribution, *see* Distributions, Poisson distribution
- Poisson point processes, spatial, 55
 intensity function, 55
 emission density, 55
- Poisson process
 AIDS modeling, in, 294
- Poisson regression, xxii, 55–56
- Poor man's data augmentation algorithm
 (PMDA), 229–230
 imputation step in, 229
 posterior step in, 229
- Positron annihilation, 55
- Positron emission tomography, *see* PET
- Posterior
 density, 26, 213, 219, 228–229, 231–232, 249, 256
 complete-data, 229
 distribution, 3, 59, 227–231, 233, 235, 237, 240–241, 250, 253, 257–260, 264, 276
 intractable, 250, 300
 marginal density, 231
 mode, 3, 188, 220, 228, 230–233, 260
 predictive, 265, 276
 probability, 16–17, 64, 140, 206, 210, 227, 302
 step (P-step), 229–230
- Principal components, 196–197
 with missing values, 197
 regularized EM for, 197
- Prior
 conjugate, 60
- density, 26–27, 102, 229, 231, 249
 Dirichlet distribution as, 250, 260
 distribution, 27, 188, 197, 213, 215, 230, 233, 241, 245, 259, 276
 improper, 262
- Projection-Solution algorithm, 269–272
- Proportional hazards mixed effects models
 (PHEM), 191, 224
- Q*-function, *see* E-step, *Q*-function
- Quadrature vs Monte Carlo methods, 220
- Quasi-likelihood, 269–271, 273
- Quasi-Newton methods, *see* Newton-type methods, quasi-Newton methods
- Quasi-score, 36, 269–2732
 incomplete data problem, for an, 269
- Radius of convergence, 141
- Random effects model, xxii, 2, 182–183, 186–192, 202, 224–225, 242, 247
- Rao-Blackwellized estimate
 with Gibbs sampler, 243, 252–253, 261–262
- Rate of convergence, 33, 77, 111, 124, 150, 212–213, 216, 310
 censored exponential survival times example, 22
 comparison of EM, ECM, and ECME algorithms, 196
 componentwise, 100–101
 Jacobian calculation, in, 122
 ECM algorithm, of, 159, 163–164
 ECME algorithm, of, 159, 181–182
 EM algorithm, of, 99–103
 efficient data augmentation, 197–198
 EM gradient algorithm, of, 150
 fraction of missing information, relation to, 22, 33, 77, 101, 121, 142, 198, 212
 Gibbs sampler, of, 257–258
 geometric, 242
 global, 100–101
 multinomial example, 13
 rate matrix, 99–103
 linear convergence, for, 99–100
 information matrices, in terms of, 77, 101–103
 MAP estimate, in, 102
t-distribution example
 comparison of EM and ECME, 181–182
- Regression, 290
- Rejection sampling, 37, 191, 220, 224, 233–236

- Adaptive rejection sampling (ARS), 234
 majorizing constant, 233–234
 proposal density, 233–234
 SIR, comparison with, 235
 target density, 233
- Reliability analysis, 20
- Repeated measures, *see also* Variance components analysis
 AEM algorithm, 146
- Response mechanism, 266
- Restricted maximum likelihood (REML) 187–192
 Bayesian formulation of, 188, 191
 of covariance components in longitudinal data analysis, 192
- Richardson-Lucy algorithm, 57–58
- Ridge regression, 197, 294
 roughness penalty in, 28
- Right censored samples, 173
- Robust analysis, 59, 61, 207
- Robust estimation, *see* Estimation, robust
- SAGE algorithm, 35, 160, 203, 263
- Sampling-importance resampling (SIR), 230, 235
 PET, in, 230
 scaled inverted gamma distribution, 259
- Saturn image reconstruction, 57–58
- Scalable EM, 217
- Score function in GLM, 271
- Score statistic
 complete-data score statistic, 34, 38, 95, 114
 definition, 4
 incomplete-data score statistic, 34–35, 38, 70, 95, 116, 119, 138, 140, 271
 individual observations, for, 7, 116, 140
 truncation case, 116
- Scoring method, *see* Fisher’s scoring method
- Secant approximation (update)
 low rank, 6–7
 symmetric, rank-one, 152
- Self-consistency, *see* EM algorithm, self-consistency
- SEM algorithm, *see* Supplemented EM algorithm
- Shepherd’s correction, 66
- Simulate-Update algorithm, 275–276
- Simulation
 experiment for hybrid models, 149
 simulation-based methods, 2
- Simulated annealing, xxiv, 29, 228, 269, 284–285, 308
- Single photon emission computed tomography, *see* SPECT
- Smoothing step, 213
- Space-filling condition
 ECM algorithm, for, 161
- Sparse EM algorithm (SPEM) 216
- Sparse IEM algorithm (SPIEM), 217
- SPECT, 54–58
 AECM algorithm for, 204
 EMS algorithm for, 213
 LININPOS problem, as, 217–218
 OSL algorithm for, 213
- Speech recognition
 hidden Markov models for, 32, 290–291
- Speed of convergence, 20, 35, 99–100, 106, 150, 159, 198–199, 201, 212
- ECM algorithm, of, 162–163, 165, 175
- Standard error estimation, xxiii, 105, 109, 111, 120. *See also* Covariance matrix of MLE
- bootstrap approach, to, 130–131
 grouped and truncated data, for, 72–73
 Louis’ method, 132–135
 Oakes’ formula, 133–134
 OSL algorithm, for, 213–214
 penalized algorithm, for, 214–216
- Statistical computing, xx, xxii
- Stochastic EM algorithm, 36–37, 227–228
- Stochastic E-step, *see* E-step, stochastic
- Stopping criterion
 Aitken acceleration-based, 142–143
 simulated annealing, for, 285
- Sufficient statistic, *see* Exponential family, sufficient statistics
- Sundberg formulas, *see* Exponential family, regular, Sundberg formulas, in
- Supplemented EM algorithm, xxiii, 34, 120–130, 214–215, 294
- bivariate normal distribution, for missing values, with, 128–130
 monitoring convergence in, 124–125
 numerical methods in, 122–123
 stability in, 123–124
- Survival analysis, 20–24, 34, 289
 EM algorithm in, 173–175
 ECM algorithm for mixture models in, 169–173
- Survival function, counting process derivation of, 173
- t*-distribution, *see also* Distributions, *t*-distribution
 factor analyzers, 207–211

- maximum likelihood estimation of, 198–203
 - multivariate
 - known d.f., 58–62
 - unknown d.f., 176–182
 - univariate, 88, 90
- Tomography**
- Positron emission tomography, *see* PET
 - Single photon emission computed tomography, *see* SPECT
- Truncated distributions**, 2, 289
- Uniform distribution**, *see* Distributions, uniform distribution
- Univariate contaminated normal model**, 125–128
- Variance components analysis**, xxiii, 31, 37, 102, 182–192, 202, 224, 289
- ECME algorithm for, 185–187
 - Gibbs sampler analog of EM algorithm for, 262
 - Henderson's algorithms
 - EM algorithm, as, 31
 - Variational Bayesian EM algorithm (VBEM), 276–277
 - EM for MAP as special case of, 277
 - posterior predictive distribution, 276
- Wishart distribution**, *see* Distributions, Wishart distribution
- Working parameter**, *see* Data Augmentation, efficient, working parameter
- Wu's conditions**, 80–85, 161

This Page Intentionally Left Blank

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice,
Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith,
Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall,
Jozef L. Teugels*

The *Wiley Series in Probability and Statistics* is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

- † ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Second Edition*
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
- BELSLY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
- † BELSLY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
- BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Third Edition*
- BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
- BERNARDO and SMITH · Bayesian Theory
- BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
- BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
- BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
- BILLINGSLEY · Probability and Measure, *Third Edition*
- BIRKES and DODGE · Alternative Methods of Regression
- BISWAS, DATTA, FINE, and SEGAL · Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics
- BLISCHKE AND MURTHY (editors) · Case Studies in Reliability and Maintenance
- BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization
- BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
- BOLLEN · Structural Equations with Latent Variables
- BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective
- BOROVKOV · Ergodicity and Stability of Stochastic Processes
- BOULEAU · Numerical Methods for Stochastic Processes
- BOX · Bayesian Inference in Statistical Analysis
- BOX · R. A. Fisher, the Life of a Scientist
- BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*
- * BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
- BOX and FRIENDS · Improving Almost Anything, *Revised Edition*
- BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, *Second Edition*
- BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
- BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
- † BROWN and HOLLANDER · Statistics: A Biomedical Introduction
- BRUNNER, DOMHOF, and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
- BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
- CAIROLI and DALANG · Sequential Stochastic Optimization
- CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
- CHAN · Time Series: Applications to Finance
- CHARALAMBIDES · Combinatorial Methods in Discrete Distributions
- CHATTERJEE and HADI · Regression Analysis by Example, *Fourth Edition*
- CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
- CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*
- CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
- CHILÉS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
- CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*
- CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
- * COCHRAN and COX · Experimental Designs, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- CONGDON · Applied Bayesian Modelling
 CONGDON · Bayesian Models for Categorical Data
 CONGDON · Bayesian Statistical Modelling
 CONOVER · Practical Nonparametric Statistics, *Third Edition*
 COOK · Regression Graphics
 COOK and WEISBERG · Applied Regression Including Computing and Graphics
 COOK and WEISBERG · An Introduction to Regression Graphics
 CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
 COVER and THOMAS · Elements of Information Theory
 COX · A Handbook of Introductory Statistical Methods
 * COX · Planning of Experiments
 CRESSIE · Statistics for Spatial Data, *Revised Edition*
 CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
 DANIEL · Applications of Statistics to Industrial Experimentation
 DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
 * DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
 DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
 DAVID and NAGARAJA · Order Statistics, *Third Edition*
 * DEGROOT, FIENBERG, and KADANE · Statistics and the Law
 DEL CASTILLO · Statistical Process Adjustment for Quality Control
 DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
 DEMIDENKO · Mixed Models: Theory and Applications
 DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
 DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
 DEY and MUKERJEE · Fractional Factorial Plans
 DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
 DODGE · Alternative Methods of Regression
 * DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
 * DOOB · Stochastic Processes
 DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
 DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
 DRYDEN and MARDIA · Statistical Shape Analysis
 DUDEWICZ and MISHRA · Modern Mathematical Statistics
 DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
 DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
 EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
 * ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
 ENDERS · Applied Econometric Time Series
 † ETHIER and KURTZ · Markov Processes: Characterization and Convergence
 EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
 FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition*, Revised; Volume II, *Second Edition*
 FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
 FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis
 * FLEISS · The Design and Analysis of Clinical Experiments
 FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- † FLEMING and HARRINGTON · Counting Processes and Survival Analysis
- FULLER · Introduction to Statistical Time Series, *Second Edition*
- † FULLER · Measurement Error Models
- GALLANT · Nonlinear Statistical Models
- GEISSER · Modes of Parametric Statistical Inference
- GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives
- GEWEKE · Contemporary Bayesian Econometrics and Statistics
- GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
- GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
- GIFI · Nonlinear Multivariate Analysis
- GIVENS and HOETING · Computational Statistics
- GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
- GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
- GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
- GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
- GROSS and HARRIS · Fundamentals of Queueing Theory, *Third Edition*
- * HAHN and SHAPIRO · Statistical Models in Engineering
- HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
- HALD · A History of Probability and Statistics and their Applications Before 1750
- HALD · A History of Mathematical Statistics from 1750 to 1930
- † HAMPEL · Robust Statistics: The Approach Based on Influence Functions
- HANNAN and DEISTLER · The Statistical Theory of Linear Systems
- HEIBERGER · Computation for the Analysis of Designed Experiments
- HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
- HEDEKER and GIBBONS · Longitudinal Data Analysis
- HELLER · MACSYMA for Statisticians
- HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*
- HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2: Advanced Experimental Design
- HOAGLIN, MOSTELLER, and TUKEY · Exploratory Approach to Analysis of Variance
- * HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
- * HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
- HOCHBERG and TAMHANE · Multiple Comparison Procedures
- HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*
- HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
- HOGG and KLUGMAN · Loss Distributions
- HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
- HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
- HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*
- † HUBER · Robust Statistics
- HUBERTY · Applied Discriminant Analysis
- HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis, *Second Edition*
- HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory and Practice
- HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—
with Commentary
- HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data
- IMAN and CONOVER · A Modern Approach to Statistics
- † JACKSON · A User's Guide to Principle Components
- JOHN · Statistical Methods in Engineering and Quality Assurance
- JOHNSON · Multivariate Statistical Simulation
- JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
- JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, *Fifth Edition*
- JOHNSON and KOTZ · Distributions in Statistics
- JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
- JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions,
Volume 1, *Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions,
Volume 2, *Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
- JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*
- JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*
- JUREČKOVÁ and SEN · Robust Statistical Procedures: Asymptotics and Interrelations
- JUREK and MASON · Operator-Limit Distributions in Probability Theory
- KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
- KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
- KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
- KARIYA and KURATA · Generalized Least Squares
- KASS and VOS · Geometrical Foundations of Asymptotic Inference
- † KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
- KEDEM and FOKIANOS · Regression Models for Time Series Analysis
- KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
- KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
- KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
- KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
- KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions,
Second Edition
- KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models:
From Data to Decisions, *Second Edition*
- KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions,
Volume 1, *Second Edition*
- KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
- KOWALSKI and TU · Modern Applied U-Statistics
- KROONENBERG · Applied Multiway Data Analysis
- KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering
- LACHIN · Biostatistical Methods: The Assessment of Relative Risks
- LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
- LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
- LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
- LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
- LAWSON · Statistical Methods in Spatial Epidemiology
- LE · Applied Categorical Data Analysis
- LE · Applied Survival Analysis
- LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*
- LePAGE and BILLARD · Exploring the Limits of Bootstrap
- LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
- LIAO · Statistical Group Comparison
- LINDVALL · Lectures on the Coupling Method
- LIN · Introductory Stochastic Analysis for Finance and Insurance
- LINHART and ZUCCHINI · Model Selection
- LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
- LLOYD · The Statistical Analysis of Categorical Data
- LOWEN and TEICH · Fractal-Based Point Processes
- MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
- MALLER and ZHOU · Survival Analysis with Long Term Survivors
- MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
- MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
- MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
- MARCHETTE · Random Graphs for Statistical Pattern Recognition
- MARDIA and JUPP · Directional Statistics
- MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
- McCULLOCH and SEARLE · Generalized, Linear, and Mixed Models
- McFADDEN · Management of Data in Clinical Trials, *Second Edition*
- * McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
- McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data
- McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, *Second Edition*
- McLACHLAN and PEEL · Finite Mixture Models
- MCNEIL · Epidemiological Research Methods
- MEEKER and ESCOBAR · Statistical Methods for Reliability Data
- MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
- MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, *Third Edition*
- * MILLER · Survival Analysis, *Second Edition*
- MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis and Forecasting
- MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Fourth Edition*
- MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
- MUIRHEAD · Aspects of Multivariate Statistical Theory
- MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks
- MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization
- MURTHY, XIE, and JIANG · Weibull Models
- MYERS and MONTGOMERY · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- MYERS, MONTGOMERY, and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences
- † NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
- † NELSON · Applied Life Data Analysis
- NEWMAN · Biostatistical Methods in Epidemiology
- OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
- OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
- OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
- PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions
- PANJER · Operational Risk: Modeling and Analytics
- PANKRATZ · Forecasting with Dynamic Regression Models
- PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
- * PARZEN · Modern Probability Theory and Its Applications
- PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
- PIANTADOSI · Clinical Trials: A Methodologic Perspective
- PORT · Theoretical Probability for Applications
- POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
- POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality
- PRESS · Bayesian Statistics: Principles, Models, and Applications
- PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
- PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
- PUKELSHEIM · Optimal Experimental Design
- PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
- † PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
- QIU · Image Processing and Jump Regression Analysis
- * RAO · Linear Statistical Inference and Its Applications, *Second Edition*
- RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
- RENCHER · Linear Models in Statistics
- RENCHER · Methods of Multivariate Analysis, *Second Edition*
- RENCHER · Multivariate Statistical Inference with Applications
- * RIPLEY · Spatial Statistics
- * RIPLEY · Stochastic Simulation
- ROBINSON · Practical Strategies for Experimenting
- ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
- ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
- ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
- ROSS · Introduction to Probability and Statistics for Engineers and Scientists
- ROSSI, ALLENBY, and McCULLOCH · Bayesian Statistics and Marketing
- † ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
- * RUBIN · Multiple Imputation for Nonresponse in Surveys
- RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*
- RUBINSTEIN and MELAMED · Modern Simulation and Modeling
- RYAN · Modern Engineering Statistics
- RYAN · Modern Experimental Design
- RYAN · Modern Regression Methods
- RYAN · Statistical Methods for Quality Improvement, *Second Edition*
- SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
- * SCHEFFE · The Analysis of Variance
- SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- SCHOTT · Matrix Analysis for Statistics, *Second Edition*
- SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
- SCHUSS · Theory and Applications of Stochastic Differential Equations
- SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
- † SEARLE · Linear Models for Unbalanced Data
- † SEARLE · Matrix Algebra Useful for Statistics
- † SEARLE, CASELLA, and MCCULLOCH · Variance Components
- SEARLE and WILLETT · Matrix Algebra for Applied Economics
- SEBER · A Matrix Handbook For Statisticians
- † SEBER · Multivariate Observations
- SEBER and LEE · Linear Regression Analysis, *Second Edition*
- † SEBER and WILD · Nonlinear Regression
- SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
- * SERFLING · Approximation Theorems of Mathematical Statistics
- SHAFER and VOVK · Probability and Finance: It's Only a Game!
- SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions
- SMALL and MCLEISH · Hilbert Space Methods in Probability and Statistical Inference
- SRIVASTAVA · Methods of Multivariate Statistics
- STAPLETON · Linear Statistical Models
- STAPLETON · Models for Probability and Statistical Inference: Theory and Applications
- STAUDTE and SHEATHER · Robust Estimation and Testing
- STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
- STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
- STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods
- STYAN · The Collected Papers of T. W. Anderson: 1943–1985
- SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
- TAKEZAWA · Introduction to Nonparametric Regression
- TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
- THOMPSON · Empirical Model Building
- THOMPSON · Sampling, *Second Edition*
- THOMPSON · Simulation: A Modeler's Approach
- THOMPSON and SEBER · Adaptive Sampling
- THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
- TCIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness
- TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
- TSAY · Analysis of Financial Time Series, *Second Edition*
- UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
- VAN BELLE · Statistical Rules of Thumb
- VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*
- VESTRUP · The Theory of Measures and Integration
- VIDAKOVIC · Statistical Modeling by Wavelets
- VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
- WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- WEERAHANDI · Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models
- WEISBERG · Applied Linear Regression, *Third Edition*
- WELSH · Aspects of Statistical Inference
- WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
- WHITTAKER · Graphical Models in Applied Multivariate Statistics
- WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
- WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
- WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
- WOODWORTH · Biostatistics: A Bayesian Introduction
- WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
- WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization
- WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis
- YANG · The Construction Theory of Denumerable Markov Processes
- YOUNG, VALERO-MORA, and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics
- ZELTERMAN · Discrete Distributions—Applications in the Health Sciences
- * ZELLNER · An Introduction to Bayesian Inference in Econometrics
- ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.