Machine Learning: A Probabilistic Perspective Solution Manual Version 2.1

Fangqi Li,
Shanghai Jiao Tong University,
P. R. China.

Contents

1 Preface

1.1 The Second Edition

The tide of artificial intelligence (AI) has swept and reformed so many disciplines and pushed forward the borderline of the state-of-the-art. Such a situation has resulted in the positive feedback that drives even more attention and effort into the study and research of AI, together with more unsettled regret and pities.

I have participated in this study, with equal passion that any student who has not formed an exclusive view of the world should have when engaging in a booming subject.

It has been three years and four months since I started the first edition of this solution manual. Although I have received no patron and have no intention of finding one, I received several grateful and advisory messages, from which I felt more pleased than having any of my technical papers published online. For the convenience of them, I would gladly edit this manuscript again, be the time goes back to 2017. In the second edition, I tried to be more concrete in deduction so readers can follow up easily. More graphical or numerical examples were provided to increase the overall readability. At the beginning part of each chapter/after some exercises, I left some remarks, which I thought could help.

The purpose of this manuscript is, as its first edition, to complete the textbook *Machine Learning, A Probabilistic Perspective* as a closed collection of knowledge as far as I could, and to save those who lose themselves in the ocean of deduction and symbols in ML, whom any talent mind could have become for some times in his/her course with this textbook. I hope that this manuscript can help, be it ever so little, to any reader who purposedly or accidentally finds it.

I take responsibility for any typo or mistake in β-reductions.

Fangqi Li,
Shanghai Jiao Tong University,
Shanghai, P.R.China.

January the 4th, 2021.

Contact me by: solour_lfq@sjtu.edu.cn or 1524587011@qq.com.
My homepage is at: https://solour-lfq.github.io/.

第二版序

这篇文档的主体由英文书写，这一方面是因为英文是学术上最泛用的语言，有利于本文档的传播；另一方面是我认为有能力阅读 MLaPP 原教材的中国学生、汉语母语学生基本上也能畅通无阻地阅读本文档中的英文。

希望中国学者和其他可以以中文为语言写作的学者一同努力，提升中文期刊、会议的质量和中文区科研院所的硬实力、影响力，让越来越多的学者乐于用中文叙述自己的观点。

第二版修订了第一版的一些文本问题，补充了第一版在习题推理上比较缺少的理念连接，同时增加了一些示例以提升可读性。文档内的所有排版错误、推导错误由我一人负责。

李方圻

上海交通大学

2021 年 1 月 4 日

邮箱：solour_lfq@sjtu.edu.cn，1524587011@qq.com。

主页：https://solour-lfq.github.io/。

1.2 The First Edition

This document provides detailed solutions to almost all exercises in the textbook MLaPP from Chapter One to Chapter Twenty-one. A reader is assumed to find support from this document when he/she is teaching himself/herself an introductory or advanced course with MLaPP.

There are two classes for problems in MLaPP: theoretical ones and practical ones. We provide the solution to most theoretical problems. Practical problems, which base on a Matlab toolbox, are beyond the scope of this document.

I started reading MLaPP after selecting a machine learning course, but I failed to find any free compiled solution manuals. Although several publicly available projects have started working on it, the velocity has been too slow. In the end, I hope that readers can provide comments and opinions. Apart from correcting the wrong answers, those who good at using MATLAB, Latex typesetting or those who are willing to participate in the improvement of the document are always welcome to contact me.

22/10/2017

Fangqi Li
Munich, Germany
solour_lfq@sjtu.edu.cn
ge72bug@tum.de

1.3 Updating log

22/10/2017 First Chinese compilation.

02/03/2018 English version.

06/01/2020 The second edition begins.

2 Probability

The probability theory for ML is usually a small subset of elementary probability. More involved topics in probability beginning from Kolmogorov's theory to the martingale and Markov process is usually beyond the scope of an ordinary statistical ML textbook. Readers are encouraged to refer to Shiryaev's *Probability, 3rd edition*, whose first chapter gives a comprehensive summary of elementary probability theory. For supplementary materials, readers can refer to Thomas's *Elements of Information Theory* for a solid introduction.

2.1 Probability are sensitive to the form of the question that was used to generate the answer

Denote two children by A and B. The space of all experiment results, Ω is composed of:

$$\omega_1 : A \text{ is a girl, }, B \text{ is a girl,}$$

$$\omega_2 : A \text{ is a boy, }, B \text{ is a boy,}$$

$$\omega_3 : A \text{ is a girl, }, B \text{ is a boy,}$$

$$\omega_4 : A \text{ is a boy, }, B \text{ is a girl.}$$

With uniform probability measure. Denote the σ-algebra on Ω as 2^{Ω}.

In question (a), with the knowledge *there is at least one boy*, Ω is modified into:

$$\Omega' = \{\omega_2, \omega_3, \omega_4\}.$$

The event that one child is a girl is $\{\omega_3, \omega_4\}$, whose probability is:

$$\frac{|\{\omega_3, \omega_4\}|}{|\Omega'|} = \frac{2}{3}.$$

In question (b), with the knowledge that A is a boy, the reduced experiment space is:

$$\Omega'' = \{\omega_2, \omega_4\}.$$

Then the probability that B is a girl is:

$$\frac{|\{\omega_4\}|}{|\Omega''|} = \frac{1}{2}.$$

The difference in the form of the question is reflected in that Ω is reduced to different forms and the desired events vary with our questions.

2.2 Legal reasoning

Given the assertation that the criminal has the special blood type, the space of all possibilities contains $800,000 \times \frac{1}{100} = 8,000$ samples. A sample $\omega_i \in \Omega$ denotes that the i-th person with this blood type committed the crime. Let the suspect be the j-th person with this special blood type, the event that he/she was the criminal is

$$\frac{1}{|\Omega|} = \frac{1}{8,000}.$$

For question (a): The probability that an innocent person has this blood type is almost 1%, whose opposite event is *the probability that an innocent person has another blood type*, which is 99%. This event is different from *the suspect is the crimimal*.

For question (b): Justice is not measured by probability. More evidence from forensics might increase this probability to unity or reduce it to zero.

2.3 Variance of a sum

Calculate this straightforwardly:

$$\begin{aligned}\mathrm{var}[X+Y] =& \mathbb{E}[(X+Y)^2] - \mathbb{E}^2[X+Y] \\ =& \mathbb{E}[X^2] - \mathbb{E}^2[X] + \mathbb{E}[Y^2] - \mathbb{E}^2[Y] + 2\mathbb{E}[XY] - 2\mathbb{E}[X][Y] \\ =& \mathrm{var}[X] + \mathrm{var}[Y] + 2\mathrm{cov}[X,Y].\end{aligned}$$

Using the definition of operators var, cov and the linearity of expectation should yield this result easily.

2.4 Bayes rule for medical diagnosis

Let ill and positive denote the event that you are infected and are tested positive for this disease respectively. Let health denote the opposite event

of ill. Apply Bayes's rules:

$$\begin{aligned}\Pr(\text{ill}|\text{positive}) =& \frac{\Pr(\text{ill}, \text{positive})}{\Pr(\text{positive})} \\ =& \frac{\Pr(\text{ill})\Pr(\text{positive}|\text{ill})}{\Pr(\text{ill})\Pr(\text{positive}|\text{ill})+\Pr(\text{health})\Pr(\text{positive}|\text{health})} \\ =& 0.0098\end{aligned}$$

2.5 The Monty Hall problem(The dilemma of three doors)

The answer is (b). Use prize_i, choose_i, open_i to denote the event that the prize is in/the player chooses/the host opens the i-th box. Apply Bayes's rules:

$$\begin{aligned}\Pr(\text{prize}_1|\text{choose}_1, \text{open}_3) =& \frac{\Pr(\text{choose}_1)\Pr(\text{prize}_1)\Pr(\text{choose}_3|\text{prize}_1, \text{choose}_1)}{\Pr(\text{choose}_1)\Pr(\text{open}_3|\text{choose}_1)} \\ =& \frac{\Pr(\text{prize}_1)\Pr(\text{choose}_3|\text{prize}_1, \text{choose}_1)}{\Pr(\text{open}_3|\text{choose}_1)} \\ =& \frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \frac{1}{3}\end{aligned}$$

In the last step, we summarize the potential location of the prize. This is a classical example of anti-intuition results from probability.

2.6 Conditional Independence

In question (a), we have:

$$\Pr(H|e_1, e_2) = \frac{\Pr(H)\Pr(e_1, e_2|H)}{\Pr(e_1, e_2)}$$

Thus the answer is (ii).

For question (b), we have the further decomposition:

$$\Pr(H|e_1, e_2) = \frac{\Pr(H)\Pr(e_1|H)\Pr(e_2|H)}{\Pr(e_1, e_2)}$$

So both (i) and (ii) and sufficient. Moreover, we have:

$$\begin{aligned}\Pr(e_1, e_2) =& \sum_H \Pr(e_1, e_2, H) \\ =& \sum_H \Pr(H)\Pr(e_1|H)\Pr(e_2|H)\end{aligned}$$

so (iii) is sufficint as well since we can calculate $p(e_1, e_2)$ from scratch.

2.7 Pairwise independence does not imply mutual independence

Consider three boolean variables ξ_1, ξ_2, ξ_3, ξ_1 and ξ_2 take values in 0 or 1 with equal possibility independently and $x_3 = \text{XOR}(x_1, x_2)$. It is easy to prove that x_3 is independent with x_1 or x_2, but given both x_1 and x_2, the value of x_3 is determined and thereby the mutual independence fails. For a detailed examination, denote the space of experiment outcomes by

$$\Omega = \{00, 01, 10, 11\},$$

with the first component denote the value of ξ_1, the second is for ξ_2. Then ξ_1 generates the σ-algebra:

$$\mathcal{F}_1 = \{\emptyset, \Omega, \{00, 01\}, \{10, 11\}\}.$$

ξ_2 generates:

$$\mathcal{F}_2 = \{\emptyset, \Omega, \{00, 10\}, \{01, 11\}\}.$$

ξ_3 generates:

$$\mathcal{F}_3 = \{\emptyset, \Omega, \{00, 11\}, \{01, 10\}\}.$$

One can easily check that each pair out of the triplet $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$ is a pair of independent σ-algebra, hence meets the pairwise independence.

However, each pair out of the triplet $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$ can span the entire 2^Ω. Then we would have counter examples, e.g., consider

$$A = \{00, 01\} \in \mathcal{F}_1,$$

$$B = \{11\} \in \sigma(\xi_2, \xi_3).$$

Then

$$\Pr(A \cap B) = 0 \neq \Pr(A) \cdot \Pr(B) = \frac{1}{8}.$$

Hence the mutual independence does not hold.

This example comes from cryptography. The only theoretical secure (defined by Shannon) encryption system is the one-pad cipher book. With the message denoted by m in binary code, the encryption is done by XOR m with a binary key k drawn from a cipher book. The result is ciphertext

c. From a statistical point of view, c is equally likely to be the ciphertext of any message, hence the adversary cannot break the security (since the ciphertext can be equally understood as any message, so no specific message prevails). But the encryption is a deterministic process (by using a strictly one-pad cipher book, this cipher is resistant to chosen-plaintext attack as well), so the mutual independence fails.

2.8 Conditional independence iff joint factorizes

We prove that (2.129) is tantamount to (2.130). One direction is trivial by denoting:

$$g(x,z) = p(x|z)$$

$$h(y,z) = p(y|z)$$

Conversely, we have:

$$\begin{aligned} p(x|z) =& \sum_y p(x,y|z) \\ =& \sum_y g(x,z)h(y,z) \\ =& g(x,z)\sum_y h(y,z). \end{aligned}$$

And vice versa,

$$p(y|z) = h(y,z)\sum_x g(x,z).$$

Moreover, for any z:

$$\begin{aligned} 1 =& \sum_{x,y} p(x,y|z) \\ =& (\sum_x g(x,z))(\sum_y h(y,z)) \end{aligned}$$

Thus:

$$\begin{aligned} p(x|z)p(y|z) =& g(x,z)h(y,z)(\sum_x g(x,z))(\sum_y h(y,z)) \\ =& g(x,z)h(y,z) \\ =& p(x,y|z) \end{aligned}$$

2.9 Conditional independence

For question (a), the antecedent $(X \perp W|Z, Y)$ means that $\forall x \in \sigma(X)$, $\forall w \in \sigma(W)$ and $\forall v \in \sigma(Y, Z)$ we have

$$\Pr(x \cap w|v) = \Pr(x|v) \cdot \Pr(w|v).$$

The antecedent $(X \perp Y|Z)$ can be translated to that $\forall x \in \sigma(X)$, $\forall y \in \sigma(Y)$ and $\forall z \in \sigma(Z)$,

$$\Pr(x \cap y|z) = \Pr(x|z) \cdot \Pr(y|z).$$

What we desire to obtain is $\forall x \in \sigma(X)$, $\forall w \in \sigma(W)$ and $\forall z \in \sigma(Z)$,

$$\Pr(x \cap w|z) = \Pr(x|z) \cdot \Pr(w|z).$$

This is correct by having v in the first equation taking values in $\sigma(Z)$ solely. Since $\sigma(Z) \subset \sigma(Y, Z)$.

For question (b), we have the premises: $\forall x \in \sigma(X)$, $\forall y \in \sigma(Y)$, $\forall z \in \sigma(Z)$ and $\forall w \in \sigma(W)$:

$$\Pr(x \cap y|z) = \Pr(x|z) \cdot \Pr(y|z).$$

$$\Pr(x \cap y|w) = \Pr(x|w) \cdot \Pr(y|w).$$

The desired result if $\forall v \in \sigma(Z, W)$,

$$\Pr(x \cap y|v) = \Pr(x|v) \cdot \Pr(y|v).$$

Let x and y be two disjoint events, z and w be another pair of disjoints, and none of $x \cap z, x \cap w, y \cap z, y \cap w$ is empty w.r.t. the underlying probability measure. Finally, let $v = w \cup z$. One can then check that the equation above does not hold for this setting, hence the deduction in (b) is false.

(b) is intuitively false. A straightforward example is a cryptography example: group signature with three participants. The group signature is a protocol for encryption/verification that ensures a series of security requirements including:

- Anyone participant solely cannot pass the verification.
- Two participants can pass the verification.

For example, let Alice and Bob each hold half of the secret key denoted by Z and W respectively, Y denotes the ciphertext, and X denotes the plaintext. Good group encryption would meet both antecedents in (b) but fails the conclusion. An example of a naive group signature is to use a quadratic function as the secret key:

$$f(t) = x \cdot t^2 + y \cdot t + c.$$

And provide two different points $z = (t_1, f_1)$, $w = (t_2, f_2)$ from f to Alice and Bob. Both antecedents in (b) are satisfied (by correctly translating the density) since the value of x yields no information about y. However, given both z and w then y is a determinstic function of x:

$$y = \frac{f_1 - f_2}{t_1 - t_2} - x \cdot (t_1 + t_2),$$

and the independence no longer holds. Note that a third participant is necessary to reveal the entire secret key (x, y, c). In practice, the calculation is usually done on an algebraic field/group, e.g. the elliptic curves, to deal with the problem with the density of x, y, c which is usually not uniform in the real number field.

2.10 Deriving the inverse gamma density

According to the change of variables formula:

$$p(y) = p(x)|\frac{\mathrm{d}x}{\mathrm{d}y}|.$$

We have:

$$\begin{aligned}
\mathrm{IG}(y) =& \mathrm{Ga}(x) \cdot y^{-2} \\
=& \frac{b^a}{\Gamma(a)}(\frac{1}{y})^{(a-1)+2} e^{-\frac{b}{y}} \\
=& \frac{b^a}{\Gamma(a)}(y)^{-(a+1)} e^{-\frac{b}{y}}.
\end{aligned}$$

The change of variables formula is a simplified version of the Lebesgue-Rikdon Theorem, which formally addresses the transform between probability measures defined on the same space. In the simplified version, we

take the existence of the derivative $\frac{\mathrm{d}x}{\mathrm{d}y}$ for granted. In the general case, such differential is obtained by taking the limit of simple functions that meet the dominance condition. The Lebesgue Theorem was developed to properly define the *differential of one probability measure w.r.t. another probability measure*. The derived general differential is usually denoted by $\frac{\mathrm{d}\mu_1(x)}{\mathrm{d}\mu_2(x)}$ where μ_i are probability measures.

2.11 Normalization constant for a 1D Gaussian

We have:

$$\begin{aligned} C &= \int_0^{2\pi} \int_0^{\infty} r \cdot \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \mathrm{d}r\mathrm{d}\theta \\ &= 2\pi\sigma^2 \cdot \int_0^{\infty} \exp\{-u\}\,\mathrm{d}u \\ &= 2\pi\sigma^2. \end{aligned}$$

For multivariate Gaussian, the trick is to diagonalize the covariance matrix and integrate each component independently.

2.12 Expressing mutual information in terms of entropies

We have:

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\ &= \sum_{x,y} p(x,y) \log p(x|y) - \sum_x (\sum_y p(x,y)) \log p(x) \\ &= -H(X|Y) + H(X) \end{aligned}$$

Inversing X and Y yields another formula. One can proceed to show that $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

2.13 Mutual information for correlated normals

We have:

$$\begin{aligned} I(X_1; X_2) =& H(X_1) - H(X_1|X_2) \\ =& H(X_1) + H(X_2) - H(X_1, X_2) \\ =& \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}\log(2\pi)^2\sigma^4(1-\rho^2) \\ =& -\frac{1}{2}\log(1-\rho^2) \end{aligned}$$

Here we incorporate a comprehensive deduction on (2.138) and (2.139), which shall not be taken for granted. The differential entropy for a 1D-Gaussian with density function:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}$$

is

$$\begin{aligned} & -\int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}\right)\mathrm{d}x \\ &= \ln(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2}\mathbb{E}[x^2] \\ &= \frac{1}{2}\ln(2\pi e\sigma^2). \end{aligned}$$

For the multi-dimensional case, we begin by diagonalizing the covariance matrix/decoupling the components and integrating along each independent component. Under this new set of coordinates $v_1, \cdots, v_d$, the logarithm of the density can be decomposed into

$$C + \sum_{i=1}^{d} -\frac{v_i^2}{2\sigma_i^2},$$

where σ_i^2 is the i-th diagonal component in the transformed covariance matrix. The product of all diagonal components is exactly $\det\Sigma$, hence proving (2.138).

2.14 A measure of correlation

For question (a), we only have to borrow the conclusion from Exercise 2.12.:

$$\begin{aligned} r =& 1 - \frac{H(Y|X)}{H(X)} = \frac{H(X) - H(Y|X)}{H(X)} \\ =& \frac{H(Y) - H(Y|X)}{H(X)} \\ =& \frac{I(X;Y)}{H(X)} \end{aligned}$$

For question (b), we have $0 \leq r \leq 1$ in question b for $I(X;Y) \geq 0$ and $H(X|Y) \geq 0$.

For question (c), $r = 0$ iff X and Y are independent so the distance between $p(x, y)$ and $p(x) \cdot p(y)$ is zero regarding KL-divergence.

For question (d), $r = 1$ iff X is determined by, but not necessarily equal to, Y.

2.15 MLE minimizes KL divergence to the empirical distribution

Expand the KL divergence:

$$\begin{aligned} \theta =& \arg\min_{\theta} \{\mathbb{KL}(p_{\text{emp}}||q(\theta))\} \\ =& \arg\min_{\theta} \left\{ \mathbb{E}_{p_{\text{emp}}}[\log \frac{p_{\text{emp}}}{q(\theta)}] \right\} \\ =& \arg\min_{\theta} \left\{ -H(p_{\text{emp}}) - \sum_{\mathbf{x} \in \text{dataset}} (\log q(\mathbf{x}; \theta)) \right\} \\ =& \arg\max_{\theta} \left\{ \sum_{\mathbf{x} \in \text{dataset}} \log p(\mathbf{x}; \theta) \right\} \end{aligned}$$

We use the weak law of large numbers in the third step and drop the entropy of empirical distribution, which is independent of θ, in the last step. The other direction of optimization is $\arg\min_{\theta} \{\mathbb{KL}(q(\theta)||p_{\text{emp}})\}$. It contains an expectation term w.r.t. $q(\theta)$ and is harder to solve.

2.16 Mean, mode, variance for the beta distribution

Firstly, we derive the mode for beta distribution by differentiating the pdf:

$$\frac{\mathrm{d}}{\mathrm{d}x}x^{a-1}(1-x)^{b-1} = [(1-x)(a-1)-(b-1)x]x^{a-2}(1-x)^{b-2}$$

Setting this to zero yields:

$$\text{mode} = \frac{a-1}{a+b-2}$$

Secondly, derive the moment in beta distribution:

$$\begin{aligned}\mathbb{E}[x^N] =& \frac{1}{B(a,b)}\int x^{a+N-1}(1-x)^{b-1}dx \\ =& \frac{B(a+N,b)}{B(a,b)} \\ =& \frac{\Gamma(a+N)\Gamma(b)}{\Gamma(a+N+b)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\end{aligned}$$

Setting $N = 1, 2$:

$$\mathbb{E}[x] = \frac{a}{a+b}$$

$$\mathbb{E}[x^2] = \frac{a(a+1)}{(a+b)(a+b+1)}$$

Where we have used the properties of the Gamma function. Finally:

$$\text{mean} = \mathbb{E}[x] = \frac{a}{a+b}$$

$$\text{variance} = \mathbb{E}[x^2] - \mathbb{E}^2[x] = \frac{ab}{(a+b)^2(a+b+1)}$$

2.17 Expected value of the minimum

Let m denote the location of the left most point, we have:

$$\begin{aligned}p(m > t) =& p([X > t] \text{ and } [Y > t]) \\ =& p(X > t)p(Y > t) \\ =& (1-t)^2\end{aligned}$$

Therefore:

$$\begin{aligned}\mathbb{E}[m] &= \int_0^1 t \cdot p(m = t)\mathrm{d}t \\ &= \int_0^1 p(m > t)\mathrm{d}t \\ &= \int_0^1 (1-t)^2\mathrm{d}t \\ &= \frac{1}{3}\end{aligned}$$

For the second equation, note that:

$$p(m \geq t) = \int_t^{1-t} p(m = t')\mathrm{d}t',$$

therefore

$$\int_0^1 \int_t^{1-t} p(m = t')\mathrm{d}t'\mathrm{d}t = \int_0^1 \int_0^{t'} p(m = t')\mathrm{d}t\mathrm{d}t' = \int_0^1 t \cdot p(m = t')\mathrm{d}t'.$$

There are perhaps more intuitive solutions to this problem, for example, plotting the value of X, Y and $\min(X, Y)$ into one graph: The height of the

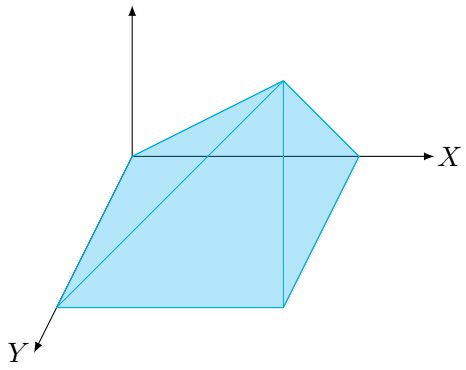


Figure. 1. Exercise 2.17, P1.

cyan pyramid at (x, y) marks the value of $\min(x, y)$, so the expectation of the statistics equals the average height, in this case also the volume of the pyramid, $\frac{1}{3}$. One can also graphically compute the average distance between X and Y from the following plot: Since the average distance between X and Y is $\frac{1}{3}$, so is that between $\min(X, Y)$ and $\max(X, Y)$. Moreover, we have $\mathbb{E}[\min(X, Y) + \max(X, Y)] = \mathbb{E}[X] + \mathbb{E}[Y] = 1$, hence $\min(X, Y) = \frac{1}{3}$.

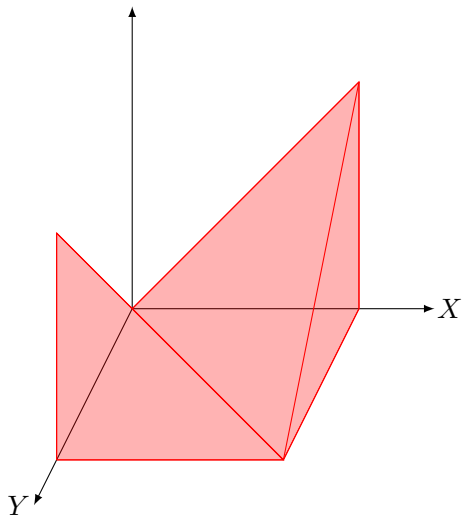


Figure. 2. Exercise 2.17, P2

However, the graphical method, although entertaining and inspiring, should not be considered as a reliable option in proving probability properties. The dependency can complicate the underlying topology (e.g., the $X-Y$ plane might have another geometry other than the Euclidean one, if X and Y are not independent), resulting in confusions and fallacies.

For example, if X is subject to a uniform distribution on $[0,1]$ while Y is uniformly distributed in $[\max(0, X-0.2), \min(1, X+0.2)]$. Then the pyramid in Fig. 1 is left with the region along the diagonal line in the X-Y plane. This generalization is insignificant since it only changes the region where the integral shall be done.

Consider the case where X and Y are independent random variables subject to truncated Gaussian centered at $\frac{1}{2}$ on $[0,1]$. The plot for visualizing $\min(X,Y)$ is the same as Fig. 1. However, to compute the expectation of $\min(X,Y)$, one cannot simply calculate the volume of the pyramid since the geometry of the X-Y plane has changed.

3 Generative models for discrete data

The Bayesian paradigm follows the following steps:

- Writing down the likelihood as a function of the parameters to be learned from the formulation of the problem.
- Choosing a corresponding prior distribution from the likelihood function such that the increment of data can be turned into easier operators.
- Writing down the posterior distribution as a function of the hyperparameters of the prior distribution and the observed data.
- If the task is to learn the (distribution of the) parameters: maximizing the posterior density at the observed data w.r.t. the hyperparameters.
- If the task is to predict: integrate out the posterior distribution conditioned on the observed data.

The Bayesian statistics beginning from this chapter provides an interesting and intuitive perspective into understanding and predicting the world. I learned Bayesian statistics from Bishop's *Pattern Recognition and Machine Learning*.

In conducting Bayesian analysis to examples provided in the exercises, we compute an extra term, the *evidence*. The evidence is the probability that a dataset being generated from a set of hyperparameters and is usually implicitly absorbed into the normalization term of the posterior distribution. The evidence is of practical value in empirical Bayesian, where we manage to select the optimal hyperparameters. In models such as variational inference, evidence plays an important role.

3.1 MLE for the Beroulli/binomial model

We begin with (3.11), which is the likelihood function of a collection of the outcomes in a coin-toss experiment $\mathcal{D}$ w.r.t. the parameter θ, the probability of heads:

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0},$$

where N_0 and N_1 are the number of tails/heads respectively.

To decompose the differential into term-independent forms, taking logarithm:

$$\ln p(\mathcal{D}|\theta) = N_1 \ln \theta + N_0 \ln(1-\theta).$$

Setting its derivative to zero:

$$\frac{\partial}{\partial \theta} \ln p(D|\theta) = \frac{N_1}{\theta} - \frac{N_0}{1-\theta} = 0,$$

yields (3.22):

$$\theta = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N},$$

where N is the size of $\mathcal{D}$.

Of course one need not turn to the logarithmic field. Differentiating $p(\mathcal{D}|\theta)$ w.r.t. θ directly gives the same result. But taking a logarithm almost always simplifies the form and the deduction procedure.

3.2 Marginal likelihood for the Beta-Bernoulli model

This exercise continues the discussion of the toy coin-toss experiment, so we borrow all symbols from the exercise above. The likelihood takes the form:

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}.$$

The prior distribution of θ takes the form:

$$p(\theta|a,b) = \mathrm{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1} = C_1(a,b) \cdot \theta^{a-1}(1-\theta)^{b-1},$$

where we adopt $C_1(a,b)$ in the hope of eliminating the ambiguity of using $\propto$, which, although simplifies the symbolization, results in countless errors.

The posterior distribution takes the form:

$$\begin{aligned} p(\theta|\mathcal{D},a,b) =& \frac{p(\theta|a,b) \cdot p(\mathcal{D}|\theta,a,b)}{p(\mathcal{D}|a,b)} \\ =& \frac{p(\theta|a,b) \cdot p(\mathcal{D}|\theta)}{p(\mathcal{D}|a,b)} \\ =& \frac{C_1(a,b)}{p(\mathcal{D}|a,b)} \cdot \theta^{N_1+a-1} \cdot (1-\theta)^{N_0+b-1}. \end{aligned}$$

The first step is the straightforward Bayesian rule, the second is the Markov property. In the last step, we adopt the equations before. Since $p(\theta|\mathcal{D}, a, b)$ should be normalized w.r.t. θ, it has to be a Beta distribution with hyperparameters $N_1 + a, N_0 + b$. We can now derive the *evidence* of $\mathcal{D}$ w.r.t. a and b explicitly. The normalization of $p(\theta|\mathcal{D}, a, b)$ indicates that:

$$\frac{C_1(a,b)}{p(\mathcal{D}|a,b)} = C_1(N_1 + a, N_0 + b),$$

so:

$$p(\mathcal{D}|a,b) = \frac{C_1(a,b)}{C_1(N_1 + a, N_0 + b)},$$

where $C_1(\cdot,\cdot)$ is the normalization factor for the Beta distribution. This is enough for deriving (3.80) by recalling the normalization of Beta distribution. The value of $p(\mathcal{D}|a,b)$ can help us select proper hyperparametes.

As for prediction:

$$\begin{aligned} p(x_{\text{new}} = 1|\mathcal{D}, a, b) &= \int p(x_{\text{new}} = 1|\theta, a, b) \cdot p(\theta|\mathcal{D}, a, b) d\theta \\ &= \int p(x_{\text{new}} = 1|\theta) \cdot p(\theta|\mathcal{D}, a, b) d\theta \\ &= \int \theta \cdot p(\theta|\mathcal{D}, a, b) d\theta \\ &= \mathbb{E}_{\text{Beta}(N_1+a, N_0+b)}(\theta) = \frac{N_1 + a}{N_1 + a + N_0 + b}. \end{aligned}$$

The first step is the Bayesian rule, the second is Markov property. The rest is straightforward algebra.

Concretely, we calcualte $p(\mathcal{D})$ where $\mathcal{D} = \{1, 0, 0, 1, 1\}$:

$$\begin{aligned} p(\mathcal{D}) =& p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)...p(x_N|x_{N-1}, x_{N-2}, ...x_1) \\ =& \frac{a}{a+b}\frac{b}{a+b+1}\frac{b+2}{a+b+2}\frac{a+1}{a+b+3}\frac{a+2}{a+b+4}. \end{aligned}$$

Rename the variables $\alpha = a + b, \alpha_1 = a, \alpha_0 = b$, we have (3.83). To derive (3.80), we make use of:

$$[(\alpha_1)..(\alpha_1 + N_1 - 1)] = \frac{(\alpha_1 + N_1 - 1)!}{(\alpha_1 - 1)!} = \frac{\Gamma(\alpha_1 + N_1)}{\Gamma(\alpha_1)}.$$

3.3 Posterior predictive for Beta-Binomial model

Straightforward algebra (recall (2.61)) gives:

$$\begin{aligned}\mathrm{Bb}(1|\alpha_1', \alpha_0', 1) =& \frac{B(\alpha_1'+1, \alpha_0')}{B(\alpha_1', \alpha_0')} \\ =& \frac{\Gamma(\alpha_0'+\alpha_1')}{\Gamma(\alpha_0'+\alpha_1'+1)} \frac{\Gamma(\alpha_1'+1)}{\Gamma(\alpha_1')} \\ =& \frac{\alpha_1'}{\alpha_1'+\alpha_0'}.\end{aligned}$$

The hint provided in the textbook is incorrect by mistaking

$$\Gamma(a) = (a-1) \cdot \Gamma(a-1)$$

for

$$\Gamma(a) = a \cdot \Gamma(a-1).$$

3.4 Beta updating from censored likelihood

The derivation is straightforward:

$$\begin{aligned}p(\theta, X<3) =& p(\theta) \cdot p(X<3|\theta) \\ =& p(\theta) \cdot \left(\sum_{i=0}^{2} p(X=i|\theta) \right) \\ =& \mathrm{Beta}(\theta|1,1) \cdot \left(\sum_{i=0}^{2} \mathrm{Bin}(i|5,\theta) \right),\end{aligned}$$

with

$$\mathrm{Bin}(m|n,\theta) = \binom{n}{m} \cdot \theta^m \cdot (1-\theta)^{n-m}$$

is the probability that m heads appear in n times of experimens with the probability of head θ. The posterior distribution over θ, in this case, becomes much more involved.

3.5 Uninformative prior for log-odds ratio

Since:

$$\phi = \log \frac{\theta}{1-\theta}.$$

By using change of variables formula:

$$p(\theta) = p(\phi) \cdot |\frac{\mathrm{d}\phi}{\mathrm{d}\theta}| \propto \frac{1}{\theta(1-\theta)},$$

hence

$$p(\theta) = \mathrm{Beta}(\theta|0,0).$$

That is to say, we can generate samples subject to a Beta distribution by transforming samples drawn from a uniform distribution. This trick is of significant practical value. Direct sampling from a Beta distribution requires inverting its cumulative probability function, which involves too much computation.

3.6 MLE for the Poisson distribution

The Poisson distribution plays a central role in the stochastic process, e.g., the queueing theory. If data are assumed to be generated from a similar process then the Bayesian analysis of the Poisson distribution derived in this exercise and the next can be applied directly. The likelihood of data for a Poisson distribution is (assuming i.i.d.):

$$p(\mathcal{D}|\lambda) = \prod_{n=1}^{N} \mathrm{Poi}(x_n|\lambda) = \exp(-\lambda N) \cdot \lambda^{\sum_{n=1}^{N} x_n} \cdot \frac{1}{\prod_{n=1}^{N} x_n!}.$$

Setting the derivative of the likelihood w.r.t. λ to zero:

$$\frac{\partial}{\partial\lambda} p(\mathcal{D}|\lambda) = \frac{\exp(-\lambda N) \cdot \lambda^{(\sum_{n=1}^{N} x_n)-1}}{\prod_{n=1}^{N} x_n!} \left\{ -N\lambda + \sum_{n=1}^{N} x_n \right\}.$$

Thus:

$$\lambda_{\mathrm{MLE}} = \frac{\sum_{n=1}^{N} x_n}{N}.$$

The formulation could be made easier by taking logarithm (since the Poisson distribution can be considered an element of the exponential family as well):

$$\log p(\mathcal{D}|\lambda) = -\lambda \cdot N + \left(\sum_{n=1}^{N} x_n \right) \cdot \log \lambda,$$

where we have omitted the term independent of λ.

3.7 Bayesian analysis of the Poisson distribution

The conjugate prior for the Poisson distribution is the Gamma distribution:

$$\mathrm{Ga}(\lambda|a,b) = \frac{b^a}{\Gamma(a)} \cdot \lambda^{a-1} \cdot \exp(-\lambda \cdot b).$$

The posterior for a Bayesian Poisson model reads:

$$\begin{aligned} p(\lambda|\mathcal{D},a,b) =& \frac{p(\mathcal{D}|\lambda) \cdot p(\lambda|a,b)}{p(\mathcal{D}|a,b)} \\ =& \frac{b^a}{\prod_{n=1}^N x_n! \cdot p(\mathcal{D}|a,b) \cdot \Gamma(a)} \cdot \lambda^{a+\sum_{n=1}^N x_n - 1} \cdot \exp\left(-\lambda \cdot (N+b)\right) \\ =& \mathrm{Ga}(a + \sum_{n=1}^N x_n, N+b), \end{aligned}$$

in which the last step follows the normalization condition. We now have the evidence:

$$p(\mathcal{D}|a,b) = \frac{b^a \cdot \Gamma(a + \sum_{n=1}^N x_n)}{\prod_{n=1}^N x_n! \cdot (N+b)^{a+\sum_{n=1}^N x_n} \cdot \Gamma(a)}.$$

Finally, be a and b approximate zero, the posterior mean approaches $\frac{\sum_{n=1}^N x_n}{N}$, the same as the MLE, i.e., $a = 0, b = 0$ is a non-informative prior. One should note that this property does not hold for all Bayesian analysis, setting all hyperparameters to zero does not necessarily gracefully degenerate the posterior mean to the MLE. Since the names and definitions of those symbols might differ.

3.8 MLE for the uniform distribution

The Bayesian analysis for the uniform distribution seems to be of less significance since uniform distribution appears to appear less frequently than other continuous distributions. But the exercises remain good introductory examples.

The likelihood for the uniform distribution is a truncated function, whose domain is $[-a, a]$, so we must have $a \geq \max_i \{|x_i| \in \mathcal{D}\}$. Then the likelihood lookes like:

$$p(\mathcal{D}|a) = \prod_{i=1}^n \frac{1}{2a},$$

or generally:

$$p(\mathcal{D}|a) = \mathbb{I}[a \geq \max_i \{|x_i| \in \mathcal{D}\}] \cdot (2a)^{-n}.$$

For question (a), in order to maximize this value with $a \geq \max_n \{|x_n| \in \mathcal{D}\}$, the outcome is:

$$a_{\text{MLE}} = \max_i \{|x_i| \in \mathcal{D}\}.$$

For question (b), if $|x_{n+1}| > \max_{i=1}^n \{|x_i|\}$ then $p(x_{n+1})$ is zero. Otherwise the probability is $\frac{1}{2 \cdot a_{\text{MLE}}}$.

For question (c), we believe that MLE for the uniform distribution is *not fluent enough* since when x_{n+1} passes $\pm \max_{i=1}^n \{|x_i|\}$, the predicted probability drops as a step function, which is undesired for a continuous distribution.

3.9 Bayesian analysis of the uniform distribution

The conjugate prior for uniform distribution is the Pareto distribution, whose density function is defined by:

$$p(\theta|K, b) = \text{Pa}(\theta|K, b) = K \cdot b^K \cdot \theta^{-(K+1)} \cdot \mathbb{I}[\theta \geq b].$$

Let $m = \max \{|x_i|\}_{i=1}^n$, the joint distribution of θ and $\mathcal{D}$ is:

$$\begin{aligned} p(\theta, \mathcal{D}|K, b) =& p(\theta|K, b) \cdot p(\mathcal{D}|\theta) \\ =& K \cdot b^K \cdot \theta^{-(K+1)} \cdot \mathbb{I}[\theta \geq b] \cdot \mathbb{I}[\theta \geq m] \cdot (\theta)^{-n} \\ =& K \cdot b^K \cdot \theta^{-(K+n+1)} \cdot \mathbb{I}[\theta \geq \max(b, m)]. \end{aligned}$$

Now $p(\theta, \mathcal{D}|K, b) = p(\mathcal{D}|K, b) \cdot p(\theta|\mathcal{D}, K, b)$, hence the posterior distribution depends on θ through:

$$\theta^{-(K+n+1)} \cdot \mathbb{I}[\theta \geq \max(b, m)].$$

So the posterior distribution is another Pareto distribution with hyperparameters $K + n, \max(b, \max \{|x_i|\}_{i=1}^n)$.

The evidence is computed from the Bayesian rule:

$$\begin{aligned} p(\mathcal{D}|K, b) &= \int_0^\infty p(\mathcal{D}, \theta|K, b) \mathrm{d}\theta \\ &= \int_{\max(b,m)}^\infty \frac{K \cdot b^K}{\theta^{K+n+1}} \mathrm{d}\theta. \end{aligned}$$

The rest is trivial calculus.

3.10 Taxicab problem

Some similar entertaining problems are *guessing the number of piano tuners from the average time for a tuner to arrive in one guest's house*, etc.

For question (a), we begin with hyperparameters $K = 0$, $b = 0$, which is improper since the Pareto distribution cannot normalize. With $\mathcal{D} = \{100\}$, we have the posterior distribution another Pareto distribution with $K = 1$ and $b = 100$, i.e.,

$$p(\theta|\mathcal{D}) = \frac{100}{\theta^2} \cdot \mathbb{I}[\theta \geq 100].$$

For question (b), we firstly derive the distribution of the taxi index:

$$\begin{aligned} p(x|\mathcal{D}, K, b) &= \int_0^\infty p(x, \theta) \mathrm{d}\theta \\ &= \int_0^\infty p(x|\theta) \cdot p(\theta|\mathcal{D}, K, b) \mathrm{d}\theta \\ &= \int_{100}^\infty \mathbb{I}[x \leq \theta] \cdot \frac{1}{\theta} \cdot \frac{100}{\theta^2} \mathrm{d}\theta \\ &= \int_{\max(x,100)}^\infty \frac{100}{\theta^3} \mathrm{d}\theta \\ &= 50 \cdot \max(x, 100)^{-2}, \end{aligned}$$

whose plots look very much similar to that of electrical potential along an axis that penetrates the center of a conductor sphere with radius 100, through declines exponentially faster.

The posterior mode of x is any number in $[0, 100]$.

The posterior mean of x is:

$$\mathbb{E}(x) = \sum_{x=0}^{100} \frac{x}{200} + \sum_{x=100}^{\infty} \frac{50}{x},$$

whose second term diverges, so the posterior mean does not exist.

The posterior median is 99.5, since:

$$\sum_{x=0}^{99} \frac{1}{200} < 0.5 < \sum_{x=0}^{100} \frac{1}{200}.$$

Question (c) is identical to (b), as we have adopted a Bayesian treatment for (b).

For question (d), we have:

$$p(x = 100|\mathcal{D}, K, b) = \frac{1}{200},$$
$$p(x = 50|\mathcal{D}, K, b) = \frac{1}{200},$$
$$p(x = 150|\mathcal{D}, K, b) = \frac{1}{450}.$$

For question (e), we might adopt better K and b with expert knowledge and collect more samples.

3.11 Bayesian analysis of the exponential distribution

The exponential distribution is also crucial for the queueing theory. The log-likelihood for an exponential distribution with density:

$$p(x|\theta) = \theta \cdot \exp(-\theta \cdot x)$$

is:

$$\ln p(\mathcal{D}|\theta) = N \cdot \ln \theta - \theta \cdot \sum_{n=1}^{N} x_n,$$

whose derivative is:

$$\frac{\partial}{\partial \theta} \ln p(\mathcal{D}|\theta) = \frac{N}{\theta} - \sum_{n=1}^{N} x_n$$

Thus for question (a), we have:

$$\theta_{\text{MLE}} = \frac{N}{\sum_{n=1}^{N} x_n}.$$

For question (b), $\theta_{\text{MLE}} = 5$.

For question (c), we begin with an exponential prior distribution:

$$p(\theta|\lambda) = \lambda \cdot \exp(-\lambda \cdot \theta),$$

whose expectation is:

$$\int_0^\infty \lambda \cdot \theta \cdot \exp(-\lambda \cdot \theta) \mathrm{d}\theta.$$

Integration by parts (or resort to the nomarlization term of the Gamma distribution) yields:

$$\mathbb{E}(\theta) = \frac{1}{\lambda}.$$

So $\hat{\lambda} = 3$.

For question (d), the posterior distribution is:

$$\begin{aligned} p(\theta|\mathcal{D}, \lambda) &= \frac{p(\mathcal{D}|\theta) \cdot p(\theta|\lambda)}{p(\mathcal{D}|\lambda)} \\ &= \frac{1}{p(\mathcal{D}|\lambda)} \cdot \theta^N \cdot \exp(-\theta \cdot \sum_{n=1}^{N} x_n) \cdot \lambda \cdot \exp(-\lambda \cdot \theta) \\ &= \frac{\theta^N \cdot \lambda}{p(\mathcal{D}|\lambda)} \cdot \exp\left(-\theta \cdot \left(\lambda + \sum_{n=1}^{N} x_n\right)\right). \end{aligned}$$

Hence the posterior is a Gamma distribution with hyperparameters:

$$a = N + 1,$$
$$b = \lambda + \sum_{n=1}^{N} x_n.$$

The evidence is given by: $\frac{\lambda \cdot \Gamma(a)}{b^a}$, a function of λ and $\mathcal{D}$. Hence the exponential distribution is not the conjugate distribution of itself, answering question (e).

For question (f), the posterior mean is the mean of the Gamma distribution:

$$\frac{a}{b} = \frac{N+1}{\lambda + \sum_{n=1}^{N} x_n}.$$

Compared with the MLE, the posterior mean has additional terms for both the numerator and the denominator as basic knowledge when N is relatively small. The influence of using this prior is tantamount to introducing a prior sample with value λ.

3.12 MAP estimation for the Bernoulli with non-conjugate priors

For question (a), we adopt the different prior:

$$p(\theta) = \begin{cases} 0.5, \text{ if } \theta = 0.5, \\ 0.5, \text{ if } \theta = 0.4, \end{cases}$$

The posterior distribution now reads:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) \cdot p(\theta)}{p(\mathcal{D})},$$

whose support is $\{0.4, 0.5\}$, so the MAP is:

$$\max_{\theta \in \{0.4,0.5\}} \left\{ \theta^{N_1} \cdot (1-\theta)^{N_0} \right\}.$$

For question (b), it is intuitive that the non-conjugate has better performance when N is small. But the conjugate Bayesian method prevails with N grows. For a solid verification, consider the case where N is large, the probability that $\frac{N_1}{N}$ deviates ϵ from 0.41 can be bounded by the Chernoff bounding. Let $\{\xi_n\}_{n=1}^N$ be a collection of i.i.d. Bernoulli random variables with distribution:

$$\xi_n = \begin{cases} 1, & \text{with probability } 0.41, \\ 0, & \text{with probability } 0.59, \end{cases}$$

Denote $X = \sum_{n=1}^N \xi_n$ as the random variable marks their summation. Then:

$$\begin{aligned} \Pr(X \geq N \cdot (0.41+\epsilon)) &= \Pr(\mathrm{e}^{\lambda \cdot X} \geq \mathrm{e}^{N\lambda(0.41+\epsilon)}) \\ &\leq \frac{\mathbb{E}[\mathrm{e}^{\lambda \cdot X}]}{\mathrm{e}^{N\lambda(0.41+\epsilon)}} \\ &= \frac{\left(\mathbb{E}[\mathrm{e}^{\lambda \cdot \xi_0}]\right)^N}{\mathrm{e}^{N\lambda(0.41+\epsilon)}} \\ &= \frac{\left(\mathbb{E}[\mathrm{e}^{\lambda \cdot \xi_0}]\right)^N}{\mathrm{e}^{N\lambda(0.41+\epsilon)}} \\ &= \frac{\left(0.41\mathrm{e}^{\lambda} + 0.59\right)^N}{\mathrm{e}^{N\lambda(0.41+\epsilon)}}. \end{aligned}$$

Where λ can be an arbitrary positive number. The probability that X exceeds $N \cdot (0.41+\epsilon)$, denoted by P_1 is bounded by the lower bound of the last line in the deduction above. With $N = 1000$, $\epsilon = 0.05$, the probability is numerically bounded by 0.00602. The other side of error $\Pr(X \leq N \cdot (0.41-\epsilon))$, whose probability is P_2, can be derived in a similar way. The probability that the Bayesian way is dominated by the non-uniform prior is no higher than $P_1 + P_2$. Taking $\epsilon \to 0.1$ and $N \to \infty$, this bound remains negligible.

3.13 Posterior predictive distribution for a batch of data with the dirichlet-multinomial model

The likelihood for Dirichlet-multinomial model is:

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{K} \theta_k^{N_k^{\text{old}}},$$

following the symbols defined in the textbook. The conjugate prior is the Dirichlet distribution:

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \cdot \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},$$

where θ is a K-dimension simplex. The (3.37) in the textbook mistake θ for $\mathbf{x}$.

The posterior distribution is another Dirichlet distribution with update:

$$\alpha_k + N_k^{\text{old}} \leftarrow \alpha_k.$$

To predict a new batch of data $\tilde{\mathcal{D}}$, we begin with one sample $x \in \tilde{\mathcal{D}}$:

$$\begin{aligned} p(x = k|\mathcal{D}, \alpha) &= \int_\theta p(x = k|\theta) \cdot p(\theta|\mathcal{D}, \alpha) \mathrm{d}\theta \\ &= \mathbb{E}_{\text{Dir}}[\theta_k], \end{aligned}$$

where the expectation is computed w.r.t. the posterior Dirichlet distribution, hence is:

$$\frac{\alpha_k + N_k^{\text{old}}}{\sum_{t=1}^{K} \alpha_t + N_t^{\text{old}}}.$$

Finally,

$$\begin{aligned} p(\tilde{\mathcal{D}}|\mathcal{D}, \alpha) &= \prod_{x \in \tilde{\mathcal{D}}} p(x|\mathcal{D}, \alpha) \\ &= \prod_{k=1}^{K} \left(\frac{\alpha_k + N_k^{\text{old}}}{\sum_{t=1}^{K} \alpha_t + N_t^{\text{old}}} \right)^{N_k^{\text{new}}}. \end{aligned}$$

3.14 Posterior predictive for Dirichlet-multinomial

For question (a). In this concrete case we have $K = 27$, $N = 2,000$, any component of α be 10, and $N_{\text{e}}^{\text{old}} = 260$. To derive $p(x_{2001} = \text{e}|\mathcal{D})$, we

resort to the deduction in Exercise 3.13:

$$p(x_{2001} = \mathrm{e}|\mathcal{D}) = \frac{10 + 260}{10 \times 27 + 2000} \approx 0.1189.$$

For question (b), the independence between characters is still ignored, bringing no significant change to the computation:

$$p(x_{2001} = \mathrm{p}, x_{2002} = \mathrm{a}|\mathcal{D}) = \frac{10 + 87}{2270} \cdot \frac{10 + 100}{2270} \approx 0.0021.$$

3.15 Setting the hyper-parameters I

Solve for:

$$\frac{\alpha_1}{\alpha_1 + \alpha_2} = m,$$
$$\frac{\alpha_1 \cdot \alpha_2}{(\alpha_1 + \alpha_2)^2 \cdot (\alpha_1 + \alpha_2 + 1)} = v.$$

We have:

$$\alpha_2 = \frac{m \cdot (1 - m)^2}{v} + m - 1,$$
$$\alpha_1 = \alpha_2 \cdot \frac{m}{1 - m}.$$

3.16 Setting the beta hyper-parameters II

```
import math
m=0.15
l=0.05
u=0.3
MC=1000
delta=(u-l)/MC
def pm(a2):
    a1=a2*m/(1-m)
    pivot=l
    mass=0
    B=math.gamma(a1)*math.gamma(a2)/math.gamma(a1+a2)
    for i in range(MC):
        pivot=pivot+delta
        mass=mass+pivot**(a1-1)*(1-pivot)**(a2-1)
    mass=mass*delta/B
    return mass
```

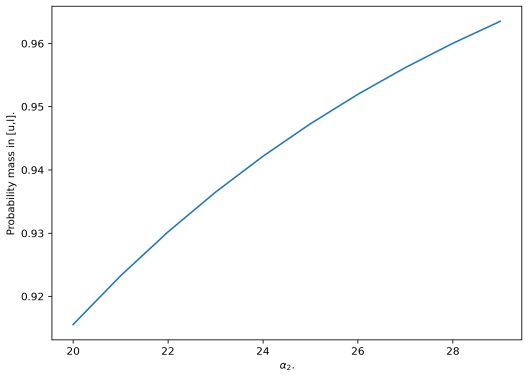


Figure. 3. Exercise 3.16.

The result of which is better demonstrated through the graph: So the optimal choice is $\alpha = 26$. This is tantamount to adopt 32 extra samples.

3.17 Marginal likelihood for beta-binomial under uniform prior

The marginal likelihood is given by:

$$p(N_1|N) = \int_0^1 p(N_1, \theta|N)\mathrm{d}\theta = \int_0^1 p(N_1|\theta, N) \cdot p(\theta)\mathrm{d}\theta.$$

Plug in:

$$p(N_1|\theta, N) = \mathrm{Bin}(N_1|\theta, N),$$

$$p(\theta) = \mathrm{Beta}(\theta|1, 1).$$

Thus:

$$\begin{aligned} p(N_1|N) =& \int_0^1 \binom{N}{N_1} \cdot \theta^{N_1} \cdot (1-\theta)^{N-N_1}\mathrm{d}\theta \\ =& \binom{N}{N_1} \cdot B(N_1 + 1, N - N_1 + 1) \\ =& \frac{N!}{N_1! \cdot (N-N_1)!} \cdot \frac{N_1! \cdot (N-N_1)!}{(N+1)!} \\ =& \frac{1}{N+1}. \end{aligned}$$

Where B is the regulizer for a Beta distribution:

$$B(a,b) = \frac{\Gamma(a)\cdot\Gamma(b)}{\Gamma(a+b)}.$$

The physics behind this setting is that if no prior information is introduced then all $N+1$ possibilities are equally likely to appear.

3.18 Bayes factor for coin tossing

The Bayes factor for hypothesis test is defined by:

$$\mathrm{BF}_{1,0} = \frac{p(\mathrm{data}|H_1)}{p(\mathrm{data}|H_0)},$$

where H_0 is the null hypothesis.

We have:

$$p(\mathrm{data}|H_0) = \mathrm{Bin}(9|0.5,10) = \binom{10}{9}\cdot 0.5^{10} \approx 0.00977.$$

And

$$p(\mathrm{data}|H_1) = \frac{1}{10+1} \approx 0.09091,$$

according to Exercise 3.17. The Bayes factor is approximately 9.3.

When $N = 100$ and $N_1 = 90$, the Bayes factor is:

$$\frac{\frac{1}{100+1}}{\binom{100}{90}\cdot 0.5^{100}} > \frac{2^{100}}{101^{11}} \approx 113622530.$$

When $\frac{N}{N_1}$ remains a constant deviated from 0.5, the larger N is, the more likely that the coin is biased. This is an intuitive conclusion from the law of large numbers.

3.19 Irrelevant features with naive Bayes

The log-likelihood is defined by:

$$\log p(\mathbf{x}_i|c,\theta) = \sum_{w=1}^{W} x_{iw}\cdot\log\frac{\theta_{cw}}{1-\theta_{cw}} + \sum_{w=1}^{W}\log(1-\theta_{cw}).$$

In a succint way:

$$\log p(\mathbf{x}_i|c,\theta) = \phi(\mathbf{x}_i)^{\mathrm{T}}\beta_c,$$

where:

$$\phi(\mathbf{x}_i) = (\mathbf{x}_i, 1)^{\mathrm{T}},$$

$$\beta_c = \left(\log \frac{\theta_{c1}}{1-\theta_{c1}}, \ldots \sum_{w=1}^{W} \log(1-\theta_{cw}) \right)^{\mathrm{T}}.$$

For question (a):

$$\begin{aligned} \log \frac{p(c=1|\mathbf{x}_i)}{p(c=2|\mathbf{x}_i)} &= \log \frac{p(c=1) \cdot p(\mathbf{x}_i|c=1)}{p(c=2) \cdot p(\mathbf{x}_i|c=2)} \\ &= \log \frac{p(\mathbf{x}_i|c=1)}{p(\mathbf{x}_i|c=2)} \\ &= \phi(\mathbf{x}_i)^{\mathrm{T}}(\beta_1 - \beta_2). \end{aligned}$$

For question (b), with:

$$\log \frac{p(c=1|\mathbf{x}_i)}{p(c=2|\mathbf{x}_i)} = \log \frac{p(c=1)}{p(c=2)} + \phi(\mathbf{x}_i)^T(\beta_1 - \beta_2),$$

a word w will not affect this posterior measure as long as:

$$x_{iw}(\beta_{1,w} - \beta_{2,w}) = 0$$

Hence if:

$$\theta_{c=1,w} = \theta_{c=2,w},$$

then it cannot affect the classification decision. That is to say, w appear in class 1 and 2 with the same frequency.

For question (c), we have:

$$\hat{\theta}_{1,w} = 1 - \frac{1}{2+N_1},$$

$$\hat{\theta}_{2,w} = 1 - \frac{1}{2+N_2}.$$

They are different when $N_1 \neq N_2$ so the bias effect remains. However, this bias reduces when N grows large.

For question (d), using information theory would be a solid option.

3.20 Class conditional densities for binary data

For question (a), we have:

$$p(\mathbf{x}|y=c) = \prod_{i=1}^{D} p(x_i|y=c, x_1, ..., x_{i-1}).$$

The number of parameter in this case is:

$$C \cdot \sum_{i=0}^{D-1} 2^i = C \cdot (2^{D+1} - 2) = \mathcal{O}(C \cdot 2^D).$$

For question (b) and (c), the overfitting is generally assumed to decline when N grows. The dependence within the variables generally hinder generalization when N is small, but it could correctly capture the dependence, be it exist.

For question (d), fitting each parameter for the naive Bayes model requires averaging all components in the samples belong to one class, which is of complexity order $\mathcal{O}(N)$. The total complexity is thus of order $\mathcal{O}(C \cdot D \cdot N)$. For the full Bayes classifier, the complexity is of order $\mathcal{O}(C \cdot 2^D \cdot N)$.

For question (e), the cost of inference in the naive Bayes model is $\mathcal{O}(C \cdot D)$ for each sample. While that for full Bayes model is $\mathcal{O}(C \cdot 2^D)$.

For question (f), we have:

$$p(y|\mathbf{x}_v, \theta) \propto p(y|\theta) \cdot p(\mathbf{x}_v|\theta) \propto \sum_{\mathbf{x}_h} p(\mathbf{x}_v, \mathbf{x}_h|\theta),$$

where we assumed a uniform prior on all classes since it is not the bottleneck of complexity. The complexity for the naive model is then:

$$\mathcal{O}(C \cdot v \cdot 2^h),$$

while that for the full model is:

$$\mathcal{O}(C \cdot 2^v) \cdot 2^h).$$

3.21 Mutual information for naive Bayes classifiers with binary features

By definition:

$$I(X;Y) = \sum_{x_j} \sum_{y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j) \cdot p(y)}.$$

For binary features, the value of x_j is either zero or one. Given $\pi_c = p(y = c), \theta_{jc} = p(x_j = 1|y = c), \theta_j = p(x_j = 1)$, we have the mutual information between x_j and Y be:

$$\begin{aligned} I_j &= \sum_c p(x_j = 1, c) \log \frac{p(x_j = 1, c)}{p(x_j = 1) \cdot p(c)} \\ &+ \sum_c p(x_j = 0, c) \log \frac{p(x_j = 0, c)}{p(x_j = 0) \cdot p(c)} \\ &= \sum_c \pi_c \theta_{jc} \log \frac{\theta_{jc}}{\theta_j} + (1 - \theta_{jc}) \pi_c \log \frac{1 - \theta_{jc}}{1 - \theta_j}, \end{aligned}$$

which ends in (3.76).

3.22 Fitting a naive Bayesian spam filter by hand

θ_{spam} is $\frac{3}{3+4}$.
$\theta_{\text{secret|spam}}$ is $\frac{2}{3}$.
$\theta_{\text{secret|non-spam}}$ is $\frac{1}{4}$.
$\theta_{\text{sports|non-spam}}$ is $\frac{1}{2}$.
$\theta_{\text{dollar|spam}}$ is $\frac{1}{3}$.

4 Gaussian models

Though being a family of models for continuous variables, traditional Gaussian models do not cast a more important role than models covered in the chapter before. Since in most scenarios, the assumption that the data are subject to a normal distribution is unrealistic.

However, Gaussian models are crucial for latent space analysis. Even for complex objects such as images or videos, the assumption that they can be encoded into features under a Gaussian distribution turns out to be reliable. Moreover, Gaussian models pave the way to general variational inference and Gaussian process, an important kernel machine.

For the reason above one cannot overestimate the significance of Gaussian models, be their history so long. Most of the mathematical difficulties with the Gaussian model have been covered in sections (marked with stars) in the textbook.

4.1 Uncorrelated does not imply independent

The mean for Y is:

$$\int_{-1}^{1} X^2 \mathrm{d}X = \mathbb{E}[X^2].$$

Calculate the covariance of X and Y:

$$\begin{aligned}\mathrm{cov}(X,Y) &= \int\int (X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y)) \cdot p(X,Y) \mathrm{d}X \mathrm{d}Y \\ &= \int_{-1}^{1} X(X^2 - \mathbb{E}[X^2]) \mathrm{d}X = 0,\end{aligned}$$

whose value is zero since we are intergrating an odd function in range $[-1, 1]$, hence:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = 0.$$

Independence is a much stronger condition than uncorrelation. The former exerts constraints on the σ-algebra that random variables generate while the latter only regulates the value of the expectation of a new random

variable. Decomposition $p(X,Y) = p(X) \cdot p(Y)$ is sufficient for reducing the covariance to zero, but not necessary.

4.2 Uncorrelated and Gaussian does not imply independent unless jointly Gaussian

For question (a). The p.d.f. for Y is:

$$p(Y \in [a, a+\mathrm{d}a]) = 0.5 \cdot p(X \in [a, a+\mathrm{d}a]) + 0.5 \cdot p(X \in [-a-\mathrm{d}a], -a) = p(X \in [a, a+\mathrm{d}a]),$$

since X is symmetric. So Y subject to a normal distribution $(0, 1)$.

For question (b), we have:

$$\begin{aligned}\mathrm{cov}(X,Y) =& \mathbb{E}(XY) - \mathbb{E}(X) - \mathbb{E}(Y) \\ =& \mathbb{E}_W(\mathbb{E}(XY|W)) - 0 \\ =& 0.5 \cdot \mathbb{E}(X^2) + 0.5 \cdot \mathbb{E}(-X^2) = 0.\end{aligned}$$

So they are uncorrelated.

To disprove dependence (in case of confusion), let:

$$a = \Phi^{-1}(\frac{1}{4}),$$

where Φ is the c.d.f of X, i.e.:

$$\int_{-\infty}^{a} \mathcal{N}(x|0,1)\mathrm{d}x = \frac{1}{4}.$$

Let $R_1 = (-\infty, a]$, $R_2 = (a, 0]$. The space of experiment results for $X \times Y$ is $\mathcal{R}^2$. Let $R_1 \times R_2$ be a Borel set in $\mathcal{R}^2$. Be X and Y independent, its probability measure should be $\frac{1}{16}$. However, when $X \in R_1$, it is impossible for Y to take a value from R_2. Hence the independency fails.

The rule of iterated expectation is but the Bayes rule:

$$\begin{aligned}\mathbb{E}[XY] &= \int_X \int_Y XY \cdot p(X,Y)\mathrm{d}X\mathrm{d}Y \\ &= \int_X \int_Y XY \cdot \left(\int_W p(X,Y,W)\mathrm{d}W \right) \mathrm{d}X\mathrm{d}Y \\ &= \int_W \left(\int_X \int_Y XY \cdot p(X,Y|W) \right) p(W)\mathrm{d}W.\end{aligned}$$

4.3 Correlation coefficient is between -1 and 1

Without loss of generality, assume $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. The statement:

$$-1 \leq \rho(X, Y) \leq 1,$$

is equal to

$$|\rho(X, Y)| \leq 1.$$

Hence we are to prove:

$$|\mathrm{cov}(X, Y)|^2 \leq \mathrm{var}(X) \cdot \mathrm{var}(Y)$$

Which can be drawn straightforwardly from Cauchy–Schwarz inequality. Let

$$g(t) = t^2 \cdot \mathrm{var}(X) + 2t \cdot \mathrm{var}(X, Y) + \mathrm{var}Y = \mathbb{E}[(tX + Y)^2] \geq 0.$$

Taking g's discriminator finishes the proof.

4.4 Correlation coefficient for linearly related variables is 1 or -1

When $Y = aX + b$:

$$\mathbb{E}(Y) = a \cdot \mathbb{E}(x) + b,$$

$$\mathrm{var}(Y) = a^2 \cdot \mathrm{var}(X).$$

Therefore:

$$\begin{aligned} cov(X, Y) =& \mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y) \\ =& a \cdot \mathbb{E}(X^2) + b \cdot \mathbb{E}(X) - a \cdot \mathbb{E}^2(X) - b \cdot \mathbb{E}(X) \\ =& a \cdot \mathrm{var}(X). \end{aligned}$$

Meanwhile:

$$\mathrm{var}(X) \cdot \mathrm{var}(Y) = a^2 \cdot \mathrm{var}(X).$$

These two are sufficient to derive:

$$\rho(X, Y) = \frac{a}{|a|}.$$

4.5 Normalization constant for a multidimensional Gaussian

Assume $\mu = \mathbf{0}$ w.l.o.g. If the covariance matrix already takes a diagonal form:

$$\Sigma = \begin{pmatrix} \lambda_1^{-1} & \cdots \\ \cdots & \cdots \\ \cdots & \lambda_d^{-1} \end{pmatrix},$$

then:

$$\begin{aligned} \int \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\Sigma\mathbf{x}\right)\mathrm{d}\mathbf{x} &= \int \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{d}\frac{x_i^2}{\lambda_i}\right)\right)\mathrm{d}\mathbf{x} \\ &= \prod_{i=1}^{d}\int \exp\left(-\frac{x_i^2}{2\cdot\lambda_i}\right)\mathrm{d}x_i \\ &= (2\pi\lambda_i)^{-\frac{d}{2}}. \end{aligned}$$

Plugging in $|\Sigma| = \prod_{i=1}^{d}\lambda_i^{-1}$ yields the desired normalization constant. In the second equation, using the distribution law (though somewhat intimidating).

For the general case, we begin by diagonalizing Σ into:

$$\Sigma = U^{\mathrm{T}}\Lambda U,$$

where Λ is a diagonal matrix with components $\lambda_1^{-1}\cdots\lambda_d^{-1}$ and U is a orthogonal matrix. The integral now becomes:

$$\int \exp\left(-\frac{1}{2}\left(U\mathbf{x}\right)^{\mathrm{T}}\Lambda\left(U\mathbf{x}\right)\right)\mathrm{d}\mathbf{x}.$$

Since $|U| = 1$ uniformly, we can directly rewrite the integral into:

$$\int \exp\left(-\frac{1}{2}\mathbf{u}^{\mathrm{T}}\Lambda\mathbf{u}\right)\mathrm{d}\mathbf{u}.$$

The rest is repeating the diagonal case.

4.6 Bivariate Gaussian

We have:

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{|\Sigma|}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma(\mathbf{x}-\mu)\right)$$
$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix}\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\right).$$

4.7 Conditioning a bivariate Gaussian

For question (a), we begin with the form from Exercise 4.6.:

$$p(x_1, x_2) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}.$$

By the Bayes rule:

$$p(x_2|x_1) = \frac{p(x_1, x_2)}{p(x_1)}.$$

So we only need to compute:

$$p(x_1) = \int p(x_1, x_2)\mathrm{d}x_2.$$

This can be done by *completing the square* inside $p(x_1, x_2)$ w.r.t. x_2:

$$2\pi\sigma_1\sigma_2\sqrt{1-\rho^2} \cdot p(x_1, x_2) = \exp\left(-\frac{1}{2(1-\rho^2)}\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right)$$
$$\cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{\rho^2(x_1-\mu_1)^2}{\sigma_1^2}\right)\right)$$
$$\cdot \exp\left(\frac{1}{2(1-\rho^2)}\frac{\rho^2(x_1-\mu_1)^2}{\sigma_1^2}\right)$$
$$= \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)$$
$$\cdot \exp\left(-\frac{1}{2\sigma_2^2(1-\rho^2)}\left((x_2-\mu_2) - \frac{\sigma_2\rho(x_1-\mu_1)}{\sigma_1}\right)^2\right).$$

Now we are ready to perform the integrating:

$$\int p(x_1, x_2)\mathrm{d}x_2 = \frac{\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \int \exp\left(-\frac{1}{2\sigma^2}(x_2-\mu)^2\right)\mathrm{d}x_2$$
$$= \frac{\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \sqrt{2\pi\sigma_2^2(1-\rho^2)},$$

where

$$\sigma^2 = \sigma_2^2(1-\rho^2),$$
$$\mu = \mu_2 + \frac{\sigma_2 \rho (x_1 - \mu_1)}{\sigma_1}.$$

Finally,

$$p(x_2|x_1) = \frac{\exp\left(\frac{(x_1-\mu_1)^2}{2\sigma^1}\right) \cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right)}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}}.$$

For question (b), we can further simplify the numerator of $p(x_2|x_1)$ into:

$$p(x_2|x_1) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}\left(\rho(x_1-\mu_1) - (x_2-\mu_2)\right)^2\right)}{\sqrt{2\pi(1-\rho^2)}}.$$

4.8 Whitening vs standardizing

Standardizing is a gold standard step for complex data, e.g., the batch normalization layer. Whitening, although fancy, is harder to carry out since it involves solving an eigen problem.

4.9 Sensor fusion with known variances in 1d

Denate the two observed datasets by $Y^{(1)}$ and $Y^{(2)}$, with size N_1, N_2, the likelihood is:

$$\begin{aligned} p(Y^{(1)}, Y^{(2)}|\mu) &= \prod_{n_1=1}^{N_1} p(Y_{n_1}^{(1)}|\mu) \cdot \prod_{n_2=1}^{N_2} p(Y_{n_2}^{(2)}|\mu) \\ &\propto \exp\left\{A \cdot \mu^2 + B \cdot \mu\right\}, \end{aligned}$$

where we have dropped terms independent from μ and used:

$$\begin{aligned} A =& -\frac{N_1}{2v_1} - \frac{N_2}{2v_2}, \\ B =& \frac{1}{v_1}\sum_{n_1=1}^{N_1} Y_{n_1}^{(1)} + \frac{1}{v_2}\sum_{n_2=1}^{N_2} Y_{n_2}^{(2)}. \end{aligned}$$

Differentiate the likelihood w.r.t. μ and set it to zero, we have:

$$\mu_{\text{MLE}} = -\frac{B}{2A}$$

The conjugate prior of this model must have a form proportional to $\exp\{A \cdot \mu^2 + B \cdot \mu\}$, so it is a normal distribution:

$$p(\mu|a, b) \propto \exp\left\{a \cdot \mu^2 + b \cdot \mu\right\}.$$

The posterior distribution is:

$$p(\mu|Y) \propto \exp\left\{(A + a) \cdot \mu^2 + (B + b) \cdot \mu\right\}.$$

Hence we have the MAP estimation:

$$\mu_{\text{MAP}} = -\frac{B + b}{2(A + a)}.$$

It is noticable that the MAP converges to ML estimation when observation times grow:

$$\mu_{\text{MAP}} \to \mu_{\text{ML}}.$$

The posterior distribution is another normal distribution, with:

$$\sigma^2_{\text{MAP}} = -\frac{1}{2(A + a)}.$$

For non-informative prior, we have $a = b = 0$ so $p(\mu|a, b)$ is uniform in the domain, then the MAP estimation is the same as MLE.

4.10 Derivation of information form formulae for marginalizing and conditioning

Plugging (4.93) and (4.95) into proven lines of (4.69) yields the information form formulae.

4.11 Derivation of the NIW posterior

We begin with the likelihood for a MVN:

$$p(\mathbf{X}|\mu, \Sigma) = (2\pi)^{-\frac{ND}{2}} |\Sigma|^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_i - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i - \mu)\right\}.$$

By (4.195), which can be proven by:

$$\begin{aligned}\sum_{n=1}^{N}(\mathbf{x}_i-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i-\mu) =& \sum_{n=1}^{N}(\bar{\mathbf{x}}-\mu+(\mathbf{x}_i-\bar{\mathbf{x}}))^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu+(\mathbf{x}_i-\bar{\mathbf{x}}))\\ =&N(\bar{\mathbf{x}}-\mu)^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu)+\sum_{n=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})\\ =&N(\bar{\mathbf{x}}-\mu)^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu)+\mathrm{tr}\left\{\Sigma^{-1}\sum_{n=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}}\right\}\\ =&N(\bar{\mathbf{x}}-\mu)^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu)+\mathrm{tr}\left\{\Sigma^{-1}\mathbf{S}_{\bar{\mathbf{x}}}\right\},\end{aligned}$$

where we have used the fact that $\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Z})=\mathrm{tr}(\mathbf{Z}\mathbf{Y}^{\mathrm{T}})$, with $\mathbf{Y}$ the shifted design matrix and $\mathbf{Z}=\Sigma^{-1}\mathbf{Y}$.

The conjugate prior for MVN's parameters (μ,Σ) is Normal-inverse-Wishart(NIW) distribution defined by:

$$\mathrm{NIW}(\mu,\Sigma|\mathbf{m}_0,k_0,v_0,\mathbf{S}_0)=\mathcal{N}(\mu|\mathbf{m}_0,\frac{1}{k_0}\Sigma)\cdot\mathrm{IW}(\Sigma|\mathbf{S}_0,v_0)$$

$$=\frac{1}{Z}|\Sigma|^{-\frac{v_0+D+2}{2}}\cdot\exp\left\{-\frac{k_0}{2}(\mu-\mathbf{m}_0)^{\mathrm{T}}\Sigma^{-1}(\mu-\mathbf{m}_0)-\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\mathbf{S}_0\right\}\right\}.$$

Hence the posterior reads (where we have omitted the condition on hyper-parameters):

$$p(\mu,\Sigma|\mathbf{X})\propto|\Sigma|^{-\frac{v_{\mathbf{X}}+D+2}{2}}\exp\left\{-\frac{k_{\mathbf{X}}}{2}(\mu-\mathbf{m}_{\mathbf{X}})^{\mathrm{T}}\Sigma^{-1}(\mu-\mathbf{m}_{\mathbf{X}})-\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\mathbf{S}_{\mathbf{X}}\right\}\right\},$$

where $v_{\mathbf{X}}$, $k_{\mathbf{X}}$, $\mathbf{m}_{\mathbf{X}}$ and $\mathbf{S}_{\mathbf{X}}$ are variables whose values are to be decided. Only terms that dependent on μ and Σ can explicitly enter the terms on the r.h.s.

Firstly, by comparing the exponential for $|\Sigma|$, we have:

$$v_{\mathbf{X}}=v_0+N.$$

Secondly, compare the coefficient for the term $\mu^{\mathrm{T}}\Sigma^{-1}\mu$ inside the exponential and we have:

$$k_{\mathbf{X}}=k_0+N.$$

Thirdly, check the coefficient for μ^{T} so we have:

$$N\Sigma^{-1}\bar{\mathbf{x}}+k_0\Sigma^{-1}\mathbf{m}_0=k_{\mathbf{X}}\Sigma^{-1}\mathbf{m}_{\mathbf{X}},$$

therefore:

$$\mathbf{m}_{\mathbf{X}} = \frac{N\bar{\mathbf{x}} + k_0\mathbf{m}_0}{k_{\mathbf{X}}}.$$

Finally, recall that for an arbitrary column vector A:

$$A^{\mathrm{T}}\Sigma^{-1}A = \mathrm{tr}(A^{\mathrm{T}}\Sigma^{-1}A) = \mathrm{tr}(\Sigma^{-1}AA^{\mathrm{T}}).$$

The terms that solely dependent on Σ^{-1} should equal to each other, so:

$$\mathrm{tr}(\Sigma^{-1}(k_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} + \mathbf{S}_0)) + \mathrm{tr}(\Sigma^{-1}(N\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}\mathbf{S}_{\bar{\mathbf{x}}})) = \mathrm{tr}(\Sigma^{-1}(k_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^{\mathrm{T}} + \mathbf{S}_{\mathbf{X}})).$$

Having arrived in:

$$N\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}} + \mathbf{S}_{\bar{\mathbf{X}}} + k_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} + \mathbf{S}_0 = k_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^{\mathrm{T}} + \mathbf{S}_{\mathbf{X}},$$

we obtain:

$$\mathbf{S}_{\mathbf{X}} = N\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}} + \mathbf{S}_{\bar{\mathbf{X}}} + k_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} + \mathbf{S}_0 - k_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^{\mathrm{T}}.$$

Recall the definition for mean we ends in (4.214) since:

$$\mathbf{S} = \sum_{n=1}^{N}\mathbf{x}_i\mathbf{x}_t^{\mathrm{T}} = \mathbf{S}_{\bar{\mathbf{X}}} + N\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}.$$

This finishes proving that the posterior distribution for MVN takes the form: $\mathrm{NIW}(\mathbf{m}_{\mathbf{X}}, k_{\mathbf{X}}, v_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}})$.

4.12 BIC for Gaussians

For question (a), recall that the maximum likelihood estimation for a MVN model is:

$$\mu_{\mathrm{MLE}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n,$$

$$\Sigma_{\mathrm{MLE}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\mathrm{MLE}})(\mathbf{x}_n - \mu_{\mathrm{MLE}})^{\mathrm{T}}.$$

So the likelihood reads:

$$\begin{aligned}
p(\mathcal{D}|\mu_{\mathrm{MLE}}, \Sigma_{\mathrm{MLE}}) &= \prod_{n=1}^{N} p(\mathbf{x}_n|\mu_{\mathrm{MLE}}, \Sigma_{\mathrm{MLE}}) \\
&= \prod_{n=1}^{N}(2\pi)^{-\frac{D}{2}} \cdot |\Sigma_{\mathrm{MLE}}|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}_n - \mu_{\mathrm{MLE}})^{\mathrm{T}}\Sigma_{\mathrm{MLE}}^{-1}(\mathbf{x}_n - \mu_{\mathrm{MLE}})\right) \\
&= (2\pi)^{-\frac{ND}{2}} \cdot |\Sigma_{\mathrm{MLE}}|^{-\frac{N}{2}} \cdot \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\mathrm{MLE}})^{\mathrm{T}}\Sigma_{\mathrm{MLE}}^{-1}(\mathbf{x}_n - \mu_{\mathrm{MLE}})\right).
\end{aligned}$$

Denote:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{x}_1 - \mu_{\text{MLE}} & \cdots & \mathbf{x}_N - \mu_{\text{MLE}} \end{pmatrix},$$

then $\Sigma_{\text{MLE}} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^{\text{T}}$, while the term in the exponential of the likelihood is:

$$-\frac{1}{2} \cdot \text{tr}(\mathbf{Y}^{\text{T}}\Sigma_{\text{MLE}}^{-1}\mathbf{Y}) = -\frac{1}{2} \cdot \text{tr}(\Sigma_{\text{MLE}}^{-1}\mathbf{Y}\mathbf{Y}^{\text{T}}) = -\frac{ND}{2}.$$

Thus the BIC is:

$$-\frac{ND}{2} \cdot \log(2\pi\text{e}) - \frac{N}{2} \cdot \log|\Sigma_{\text{MLE}}| - \frac{D + \frac{D(D+1)}{2}}{2} \cdot \log N.$$

For question (b), the fitting of a diagonal MVN model is tantamount to fitting D independent 1d Gaussian models simultaneously, thus the d-th diagnoal component of $\Sigma_{\text{MLE}}^{\text{d}}$ is:

$$\frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_{n,d}^2,$$

where we have assumed $\mathbf{x} = \mathbf{0}$ w.l.o.g. Thus the term inside the exponential of the likelihood remains $-\frac{ND}{2}$. So the BIC in this case is:

$$-\frac{ND}{2} \cdot \log(2\pi\text{e}) - \frac{N}{2} \cdot \log|\Sigma_{\text{MLE}}^{\text{d}}| - D \cdot \log N.$$

We observe that if all D components are mutually independent, i.e., Σ_{MLE} is diagonal then the BIC for diagonal MVN model is strictly larger than that for general MVN, hence the diagonal version is always preferred. In cases there exists dependence among components, the BIC for general MVN is still not necessarily larger than that of disgonal MVN. This is a reflection of the trade-off between complexity and generalization.

The Bayesian information criterion is an approximation of a model's evidence, $p(\mathcal{D})$. Let us start from:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta) \cdot p(\theta)\text{d}\theta,$$

where θ is the collection of all parameters within the current model. The trick here is to expand $\log p(\mathbf{x}|\theta)$ as a function of θ and taking approximation to the second order at θ_{MAP} so the first order gradient vanishes:

$$\log p(\mathbf{x}|\theta) \approx \log p(\mathbf{x}|\theta_0) - \frac{1}{2}(\theta - \theta_0)^{\text{T}}\mathbf{H}(\theta - \theta_0),$$

where $\mathbf{H}$ is the Hessian matrix at $\log p(\mathbf{x}|\theta_0)$. Thus we have:

$$p(\mathbf{x}|\theta) \approx p(\mathbf{x}|\theta_0) \cdot \exp\left(-\frac{1}{2}(\theta-\theta_0)^{\mathrm{T}}\mathbf{H}(\theta-\theta_0)\right).$$

We are now ready to perform the integral, with:

$$p(\mathcal{D}|\theta) = p(\mathbf{x}|\theta_0)^N \cdot \exp\left(-\frac{N}{2}(\theta-\theta_0)^{\mathrm{T}}\mathbf{H}(\theta-\theta_0)\right),$$

conducting the integral on the neighbour of $\theta_0 = \theta_{\mathrm{MAP}}$:

$$\begin{aligned}\int p(\mathcal{D}|\theta)\cdot p(\theta)\mathrm{d}\theta &\approx p(\mathcal{D}|\theta_{\mathrm{MAP}})\cdot p(\theta_{\mathrm{MAP}})\cdot \int \exp\left(-\frac{N}{2}(\theta-\theta_{\mathrm{MAP}})^{\mathrm{T}}\mathbf{H}(\theta-\theta_{\mathrm{MAP}})\right)\mathrm{d}\theta \\ &= p(\mathcal{D}|\theta_{\mathrm{MAP}})\cdot p(\theta_{\mathrm{MAP}})\cdot (2\pi)^{\frac{d}{2}}|N^{-1}\mathbf{H}^{-1}|^{\frac{1}{2}} \\ &= p(\mathcal{D}|\theta_{\mathrm{MAP}})\cdot p(\theta_{\mathrm{MAP}})\cdot (2\pi)^{\frac{d}{2}}\cdot N^{-\frac{d}{2}}\cdot|\mathbf{H}^{-1}|^{\frac{1}{2}},\end{aligned}$$

where d is the number of components in θ. Taking the logarithm of both side of the evidence yields the BIC. One can see how many compromises and assumptions have been applied in deriving an analytic form of the evidence, which is arguably the most complex variable for Bayesian analysis.

See also *PRML*, Section 4.4.

4.13 Gaussian posterior credible interval

Assume the prior distribution for an 1d normal distribution:

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2 = 9).$$

And the likelihood is:

$$p(x) = \mathcal{N}(x|\mu, \sigma^2 = 4).$$

Having observed n variables, we want that the probability measure of μ's posterior distribution is no less than 0.95 within an interval no longer than 1. The posterior for μ is:

$$\begin{aligned}p(\mu|D) \propto p(\mu)\cdot p(D|\mu) =& \mathcal{N}(\mu|\mu_0,\sigma_0^2)\cdot\prod_{i=1}^{n}\mathcal{N}(x_n|\mu,\sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}\cdot\prod_{i=1}^{n}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} \\ &= \exp\left\{(-\frac{1}{2\sigma_0^2}-\frac{n}{2\sigma^2})\mu^2+...\right\},\end{aligned}$$

where we have dropped the terms inrevelent with μ. The posterior variance of μ is determined by the coefficient of μ^2 in the exponential of the posterior distribution:

$$\sigma^2_{\text{post}} = \frac{\sigma_0^2 \sigma^2}{\sigma^2 + n\sigma_0^2}.$$

Since 0.95 of the probability mass for a normal distribution lies within -1.96σ and 1.96σ, we have:

$$n \geq 611.$$

4.14 MAP estimation for 1d Gaussians

Assume that the variance for this distribution σ^2 is known, and the mean μ is subject to a normal distribution with mean m and variance s^2. Similiar to the question before, the posterior takes the form:

$$p(\mu|X) \propto p(\mu) \cdot p(X|\mu).$$

So the posterior is another normal distribution, by comparing the coefficient for μ^2 in the exponential:

$$-\frac{1}{2s^2} - \frac{N}{2\sigma^2},$$

and that for μ:

$$\frac{m}{s^2} + \frac{\sum_{n=1}^N x_n}{\sigma^2},$$

we have the posterior mean and variance by completing the square:

$$\sigma^2_{\text{post}} = \frac{s^2\sigma^2}{\sigma^2 + Ns^2},$$

$$\mu_{\text{post}} = (\frac{m}{s^2} + \frac{\sum_{n=1}^N x_n}{\sigma^2}) \cdot \sigma^2_{\text{post}}.$$

This finishes question (a).

For question (b), we already knew that the MLE is:

$$\mu_{\text{MLE}} = \frac{\sum_{n=1}^N x_n}{N}.$$

As N increases, we have:

$$\begin{aligned}\lim_{N\to\infty} \mu_{\text{post}} &= \lim_{N\to\infty} \frac{\frac{\sigma^2}{s^2} \cdot m + \sum_{n=1}^N x_n}{\frac{\sigma^2}{s^2} + N} \\ &= \frac{\sum_{n=1}^N x_n}{N}.\end{aligned}$$

For question (c), when $s^2 \to \infty$, μ_{post} also converges to μ_{MLE} since $\frac{\sigma^2}{s^2} \to 0$.

For question (d), when $s^2 \to 0$, then $\frac{\sigma^2}{s^2} \to \infty$ and μ_{post} converges to m.

Both (c) and (d) are very intuitive. $s^2 \to \infty$ means a non-informative prior has been introduced so MAP is the same as MLE. $s^2 \to 0$ means that the knowledge that μ is close to m is very strong so that finite observations cannot modify this belief.

4.15 Sequential(recursive) updating of covariance matrix

For question (a), note that:

$$n\mathbf{C}_{n+1}-(n-1)\mathbf{C}_n = \sum_{i=1}^{n+1}(\mathbf{x}_i-\mathbf{m}_{n+1})(\mathbf{x}_i-\mathbf{m}_{n+1})^{\mathrm{T}} - \sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m}_n)(\mathbf{x}_i-\mathbf{m}_n)^{\mathrm{T}}.$$

Making use of:

$$\mathbf{m}_{n+1} = \frac{n\mathbf{m}_n + \mathbf{x}_{n+1}}{n+1},$$

we have:

$$\begin{aligned} n\mathbf{C}_{n+1} - (n-1)\mathbf{C}_n &= \mathbf{x}_{n+1}\mathbf{x}_{n+1}^{\mathrm{T}} - (n+1)\mathbf{m}_{n+1}\mathbf{m}_{n+1}^{\mathrm{T}} + n\mathbf{m}_n\mathbf{m}_n^{\mathrm{T}} \\ &= \frac{n}{n+1}(\mathbf{x}_{n+1} - \mathbf{m}_n)(\mathbf{x}_{n+1} - \mathbf{m}_n)^{\mathrm{T}}. \end{aligned}$$

For question (b), the complexity is $\mathcal{O}(d^2)$.

For question (c), plugging (4.281) directly into (4.279) yields (4.280).

For question (d), the complexity remains $\mathcal{O}(d^2)$.

4.16 Likelihood ratio for Gaussians

Consider a classifier for two classes, the generative distribution for them are two normal distributions $p(x|y = C_i) = \mathcal{N}(x|\mu_i, \Sigma_i)$, by the Bayes rule:

$$\log \frac{p(y=1|x)}{p(y=0|x)} = \log \frac{p(x|y=1)}{p(x|y=0)} + \log \frac{p(y=1)}{p(y=0)}.$$

The first term on r.h.s. is the ratio of likelihood probability.

When we have arbitrary covariance matrices:

$$\frac{p(x|y=1)}{p(x|y=0)} = \sqrt{\frac{|\Sigma_0|}{|\Sigma_1|}} \cdot \exp\left\{-\frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma_0^{-1}(x-\mu_0)\right\}.$$

As Σ_0, Σ_1 are arbitrary matrices, this formulation cannot be reduced further:

$$\log \frac{p(x|y=1)}{p(x|y=0)} = \frac{1}{2}\log\frac{|\Sigma_0|}{|\Sigma_1|} - \frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma_0^{-1}(x-\mu_0).$$

Note that the decision boundary ($\log \frac{p(x|y=1)}{p(x|y=0)} = 0$) is a quardratic surface in D-dimension space.

When both covariance matrixes are given by Σ:

$$\frac{p(x|y=1)}{p(x|y=0)} = \exp\left\{-\frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x-\mu_0)\right\},$$

so:

$$\begin{aligned}\log \frac{p(x|y=1)}{p(x|y=0)} &= -\frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x-\mu_0) \\ &= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left[(x-\mu_1)(x-\mu_1)^{\mathrm{T}} - (x-\mu_2)(x-\mu_2)^{\mathrm{T}}\right]\right).\end{aligned}$$

When Σ is a diagonal matrix, we have:

$$\begin{aligned}\log \frac{p(x|y=1)}{p(x|y=0)} &= -\frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x-\mu_0) \\ &= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left[(x-\mu_1)(x-\mu_1)^{\mathrm{T}} - (x-\mu_2)(x-\mu_2)^{\mathrm{T}}\right]\right) \\ &= \frac{1}{2}\mathrm{tr}\left(\Lambda^{-1}\Phi\right) \\ &= \frac{1}{2}\sum_{i=1}^{d}\lambda_i^{-1}\Phi_{i,i}.\end{aligned}$$

where:

$$\Phi = (x-\mu_1)(x-\mu_1)^{\mathrm{T}} - (x-\mu_2)(x-\mu_2)^{\mathrm{T}}.$$

Finally, if $\Sigma = \sigma^2 I$ then:

$$\log \frac{p(x|y=1)}{p(x|y=0)} = \frac{1}{2\sigma^2}\mathrm{tr}(\Phi).$$

Note that for the last three cases, a decision boundary is a linear plane in the space, since the quadratic term on x has been canceled in Φ.

4.17 LDA/QDA on height/weight data

...

4.18 Naive Bayes with mixed features

For question (a):

$$p(y=1|x_1=0,x_2=0) = \frac{p(y=1)\cdot p(x_1|y=1)\cdot p(x_2=0|y=1)}{p(x_1=0,x_2=0)} = \frac{0.5\cdot 0.5\cdot \frac{\exp(-0.5)}{\sqrt{2\pi}}}{p(x_1=0,x_2=0)}.$$

Similarly,

$$p(y=2|x_1=0,x_2=0) = \frac{0.25*0.5*\frac{1}{\sqrt{2\pi}}}{p(x_1=0,x_2=0)},$$

$$p(y=3|x_1=0,x_2=0) = \frac{0.25*0.5*\frac{\exp(-0.5)}{\sqrt{2\pi}}}{p(x_1=0,x_2=0)}.$$

A normalization yields the final result by eliminating $p(x_1=0,x_2=0)$.

For question (b), we have:

$$p(y=1|x_1=0) = 0.5,$$

$$p(y=2|x_1=0) = 0.25,$$

$$p(y=3|x_1=0) = 0.25,$$

since x_1 yields no more information for the classification label.

For question (c), we have:

$$p(y=1|x_2=0) \propto 0.5*\frac{\exp(-0.5)}{\sqrt{2\pi}},$$

$$p(y=2|x_2=0) \propto 0.25*\frac{1}{\sqrt{2\pi}},$$

$$p(y=3|x_2=0) \propto 0.25*\frac{\exp(-0.5)}{\sqrt{2\pi}}.$$

One can observe that unlike $p(y|x_1=0)$, $p(y|x_1=0)$ is different from the prior on labels.

4.19 Decision boundary for LDA with semi tied covariances

We begin from the Bayes rule:

$$p(y=1|\mathbf{x},\theta) \propto p(y=1|\theta)\cdot p(\mathbf{x}|y=1,\theta),$$

where we have omitted the terms independent of y. With a uniform prior on two classes:

$$p(y = 1|\mathbf{x}, \theta) \propto \mathcal{N}(\mathbf{x}|\mu_1, k\Sigma_0),$$
$$p(y = 0|\mathbf{x}, \theta) \propto \mathcal{N}(\mathbf{x}|\mu_0, \Sigma_0).$$

The decision boundary, in which we are interested, is a curve depicted by $f(\mathbf{x}) = 0$, where:

$$f(x) = \log \frac{p(y = 1|\mathbf{x}, \theta)}{p(y = 0|\mathbf{x}, \theta)}.$$

Therefore the decision boundary is:

$$\operatorname{tr}\left(\Sigma_0^{-1}\left[\Phi(\mathbf{x})\right]\right) = d \cdot \ln k,$$

where:

$$\Phi(\mathbf{x}) = \frac{(\mathbf{x} - \mu_1)(\mathbf{x} - \mu_1)^{\mathrm{T}}}{k} - (\mathbf{x} - \mu_0)(\mathbf{x} - \mu_0)^{\mathrm{T}}.$$

The decision boundary is a quardratic curve unless $k = 1$, which is geometrically very intuitive.

Let us consider a 2d case, focusing on the transformed coordinates where Σ_0 is the identity matrix. With out loss of generality, let $\mu_0 = (0, 0)$, $\mu_1 = (z, 0)$, denote the distance between a point p in this place from μ_0 and μ_1 by $a(p)$ and $b(p)$. The decision boundary is exactly:

$$\left\{p : \frac{b(p)^2}{k} - a(p)^2 = -\frac{1}{2}\ln k\right\}.$$

Plugging in the Cartesian representation $p(x, y)$, we ends up with a conical curve. The linear transform of the space would not change its conical nature.

4.20 Logistic regression vs LDA/QDA

The underlying assumptions for all four classifiers are as follows:

- GaussI assumes a covariance matrix as an identity matrix;
- GaussX has no prior assumption on the covariance matrix;
- LinLog assumes that different classes share the same covariance matrix;

- QuadLog has no prior assumption on covariance matrix, yet it assumes that all data from one class are subject to a normal distribution;

From the perspective of complexity we have the following order:

QuadLog=GaussX>LinLog>GaussI.

The MLE likelihood should follow the same order, this answers the question (a)-(d).

For question (e), the argument is untrue in general. For example, model M predicts two samples belonging to the first class with probability vectors $(0.49, 0.51)$ and $(0.99, 0.1)$. While M' outputs $(0.51, 0.49)$ and $(0.51, 0.49)$. Now M' is correct on both samples so $R(M) > R(M')$, but:

$$\frac{\log(0.49) + \log(0.99)}{2} > \frac{\log(0.51) + \log(0.51)}{2},$$

so $L(M) > L(M')$, this is sufficient for disproving the argument.

4.21 Gaussian decision boundaries

We have:

$$p(x|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left(-\frac{(x-\mu_i^2)}{2\sigma_i^2}\right),$$

so:

$$p(x|\mu_1, \sigma_1^2) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right),$$

$$p(x|\mu_2, \sigma_2^2) = \frac{1}{10^3 \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{(x-1)^2}{2 \times 10^6}\right).$$

The decision region satisfies:

$$\frac{p(x|\mu_1, \sigma_1^2)}{p(x|\mu_1, \sigma_1^2)} \geq 1,$$

this is tantamount to:

$$\frac{(x-1)^2}{10^6} - x^2 \geq -6 \cdot \ln 10.$$

Denote x_1 and x_2 as the zeros of this quardratic form, then

$$R_1 = [x_1, x_2].$$

When $\sigma^2 = \sigma^1$, R_1 is exactly $(\infty, \frac{1}{2})$.

One can solve for (a) and (b) by plugging in (4.289).

4.22 QDA with 3 classes

Solve (a) and (b) numerically:

```
import math
import numpy as np
mu1=np.array([0,0])
mu2=np.array([1,1])
mu3=np.array([1,-1])
s1=np.array([[0.7,0],[0,0.7]])
s2=np.array([[0.8,0.2],[0.2,0.8]])
s3=np.array([[0.8,0.2],[0.2,0.8]])
def p(x):
    f1=np.linalg.det(s1)**(-0.5)*math.exp(-0.5*(mu1-x)@np.
        linalg.inv(s1)@(mu1-x))
    f2=np.linalg.det(s2)**(-0.5)*math.exp(-0.5*(mu2-x)@np.
        linalg.inv(s2)@(mu2-x))
    f3=np.linalg.det(s3)**(-0.5)*math.exp(-0.5*(mu3-x)@np.
        linalg.inv(s3)@(mu3-x))
    return f1,f2,f3
x1=np.array([-0.5,0.5])
x2=np.array([0.5,0.5])
print(p(x1))
print(p(x2))
```

The outcome is:

```
(0.9995321962501862, 0.31309336105606445, 0.030361279346887253)
(0.9995321962501862, 1.005427487616282, 0.18990072283298057)
```

So the predicted labels are 1 and 2 respectively.

4.23 Scalar QDA

```
import math
hm=[67,79,71]
hf=[68,67,60]
def mu(h):
    return (h[0]+h[1]+h[2])/3
```

```
6  def sigma2(h):
7      m=mu(h)
8      return ((h[0]-m)**2+(h[1]-m)**2+(h[2]-m)**2)/3
9  # For question (a):
10 mu_m=mu(hm)
11 mu_f=mu(hf)
12 sigma2_m=sigma2(hm)
13 sigma2_f=sigma2(hf)
14 print(mu_m)
15 print(sigma2_m)
16 print(mu_f)
17 print(sigma2_f)
18 # \pi_{m}=\pi_{f}=0.5.
19 # For question (b):
20 temp_m=(2*math.pi*sigma2_m)**(-0.5)*math.exp(-(72-mu_m)**2/2/
       sigma2_m)
21 temp_f=(2*math.pi*sigma2_f)**(-0.5)*math.exp(-(72-mu_f)**2/2/
       sigma2_f)
22 print(temp_m/(temp_m+temp_f))
```

For question (c) we can use a naive Bayes model, which is tantamount to adopting diagonal covariance matrices for both classes.

5 Bayesian statistics

This Chapter is a continuation of Chapter 2. Section 5.7.1.5. sheds light on PAC theory, the fundamental theory underlying machine learning, elegant but somewhat impractical. Section 5.7.3.1. introduces some basic elements from reinforcement learning. Both branches require solid and involved theories, especially those from probability.

5.1 Proof that a mixture of conjugate priors is indeed conjugate

The mixed conjugate prior takes the form (5.68):

$$p(\theta) = \sum_k p(z = k) \cdot p(\theta | z = k),$$

where k is the index for mixed components. Each $p(\theta|z = k)$ is a conjugate prior for the model, i.e., $p(\theta|z = k)$ and $p(\theta|z = k, \mathcal{D})$ takes the same form.

For the posterior in this case:

$$\begin{aligned} p(\theta|\mathcal{D}) &= \sum_k p(\theta, z = k|\mathcal{D}) \\ &= \sum_k p(z = k|\mathcal{D}) \cdot p(\theta|z = k, \mathcal{D}). \end{aligned}$$

So the posterior is still a mixture of conjugate priors. This is exactly (5.69), finally:

$$\begin{aligned} p(z = k|\mathcal{D}) &= \frac{p(z = k, \mathcal{D})}{p(\mathcal{D})} \\ &= \frac{p(z = k) \cdot p(\mathcal{D}|p(z = k))}{\sum_{k'} p(z = k') \cdot p(\mathcal{D}|p(z = k'))}, \end{aligned}$$

from Bayes rules. The computation bottleneck in computing $p(z = k|\mathcal{D})$ is:

$$p(\mathcal{D}|z = k) = \int p(\mathcal{D}|\theta) \cdot p(\theta|z = k) \mathrm{d}\theta,$$

which is not a easy task as well.

5.2 Optimal threshold on classification probability

For question (a), the posterior loss expectation for choosing action $\hat{y}$, given x, is:

$$\begin{aligned}\rho(\hat{y}|x) &= \mathbb{E}_{p(y|x)}[L(y,\hat{y})] \\ &= p_0 \cdot L(\hat{y},0) + p_1 \cdot L(\hat{y},1) \\ &= p_0 \cdot L(\hat{y},0) + (1-p_0) \cdot L(\hat{y},1) \\ &= L(\hat{y},1) + p_0(L(\hat{y},0) - L(\hat{y},1)).\end{aligned}$$

Thus:

$$\rho(0|x) = \lambda_{01} - p_0 \cdot \lambda_{01},$$

$$\rho(1|x) = p_0 \cdot \lambda_{10}.$$

As both $\rho(0|x)$ and $\rho(1|x)$ are linear functions of p_0, hence the unique optimal threshold is where $\rho(0|x) = \rho(1|x)$, i.e.,

$$p_0 = \frac{\lambda_{01}}{\lambda_{01} + \lambda_{10}},$$

$$p_1 = \frac{\lambda_{10}}{\lambda_{01} + \lambda_{10}}.$$

For question (b), let $\lambda_{10} = 1$, $\lambda_{01} = 9$ will do.

5.3 Reject option in classifiers

For question (a), the posterior expected loss for choosing an non-reject action $\hat{y}$ given data x is:

$$\begin{aligned}\rho(\hat{y}|x) &= \sum_{y=1}^{C} p(y|x) \cdot L(\hat{y},y) \\ &= \mathbf{p}^{\mathrm{T}}\mathbf{l}(\hat{y}),\end{aligned}$$

where $\mathbf{p}$ is the column vector encoding $p(y|x)$ and $\mathbf{l}(\hat{y})$ is a column vector whose elements are λ_s except for the $\hat{y}$-th one. Thus the expected loss is $(1 - p(\hat{y}|x)) \cdot \lambda_s$ in this case, whose minimum is obtained by let

$$\hat{y} = \arg\max_y(p(y|x)).$$

For the reject option, the loss is uniform λ_s.

Thus one should choose reject or $\hat{y}$ by minimizing:

$$\min(\lambda_r, (1 - p(\hat{y}|x)) \cdot \lambda_s).$$

If

$$\lambda_r \leq (1 - p(\hat{y}|x)) \cdot \lambda_s,$$

then we readily adopt the reject option. This condition is tantamount to what is required to be prove:

$$p(\hat{y}|x) \leq 1 - \frac{\lambda_r}{\lambda_s}.$$

For question (b), the minimum of the expected loss is:

$$\lambda_s \cdot \min\left\{\frac{\lambda_r}{\lambda_s}, 1 - p(\hat{y}|x)\right\},$$

where $\hat{y}$ is the most probable class. When $\frac{\lambda_r}{\lambda_s}$ is negligable, the reject option would always be chosen. When $\frac{\lambda_r}{\lambda_s}$, the reject option would never be chosen.

5.4 More reject options

For question (a), we need to choose from:

$$(3, 10 \times (1 - 0.8) = 2),$$

hence the optimal decision is class 0.

For question (b), we need to choose from:

$$(3, 10 \times (1 - 0.6) = 4),$$

hence the optimal decision is reject.

For question (c), the threshold is where:

$$\lambda_r = (1 - p(\hat{y}|x)) \cdot \lambda_s.$$

Plugging in figures we have:

$$p(\hat{y}|x) = 0.7.$$

Thus $\theta_1 = 0.7$, $\theta_0 = 0.3$ by symmetry.

5.5 Newsvendor problem

We have:

$$\mathbb{E}(\pi|Q) = P\int_0^Q Df(D)\mathrm{d}D - CQ\int_0^Q f(D)\mathrm{d}D + (P-C)Q\int_Q^{+\infty} f(D)\mathrm{d}D.$$

Take derivative w.r.t. Q:

$$\frac{\partial}{\partial Q}\mathbb{E}(\pi|Q) = PQf(Q) - C\int_0^Q f(D)\mathrm{d}D - CQf(Q) + (P-C)\int_Q^{+\infty} f(D)\mathrm{d}D - (P-C)Qf(Q).$$

Setting it to zero by making use of $\int_0^Q f(D)\mathrm{d}D + \int_Q^{+\infty} f(D)\mathrm{d}D = 1$, we arrive in:

$$F(Q^*) = \frac{P-C}{P},$$

where:

$$F(Q) = \int_0^Q f(D)\mathrm{d}D$$

is the c.d.f. of D.

5.6 Bayes factors and ROC curves

If the prior on models are identical then both methods yield the same decision boundary.

5.7 Bayes model averaging helps predictive accuracy

Suppose the variable Δ is generated from a mixture of models, so:

$$p(\Delta|\mathcal{D}) = \sum_{m\in\mathcal{M}} p(\Delta|m,\mathcal{D})\cdot p(m|\mathcal{D}).$$

The Bayes model averaging (BMA) result is just:

$$p^{\mathrm{BMA}}(\Delta) = \sum_{m\in\mathcal{M}} p(\Delta|m,\mathcal{D})\cdot p(m|\mathcal{D}).$$

While that from an individual model $m \in \mathcal{M}$ is:

$$p^m(\Delta) = p(\Delta|m,\mathcal{D}).$$

The expected loss of the BMA result is:

$$\mathbb{E}_{p(\Delta)}[-\log p^{\mathrm{BMA}}(\Delta)],$$

while that of model m is:

$$\mathbb{E}_{p(\Delta)}[-\log p^m(\Delta)].$$

Now it is easy to see that:

$$\mathbb{E}_{p(\Delta)}[-\log p^m(\Delta)] - \mathbb{E}_{p(\Delta)}[-\log p^{\mathrm{BMA}}(\Delta)] = \mathbb{KL}(p^{\mathrm{BMA}}(\Delta)||p^m(\Delta)),$$

since the distribution on which the expectation is computed is just $p^{\mathrm{BMA}}(\Delta)$. Therefore the non-negativity of the KL divergence yields (5.127).

The conclusion from this exercise is of hardly any practical significance. Since the underlying distribution is usually intractable, even with the mixture Bayes model. Once the form of the latent distribution is revealed, it is obvious that other distributions result in higher loss.

5.8 MLE and model selection for a 2d discrete distribution

For question (a), the joint distribution $p(x, y|\theta_1, \theta_2)$ is given by:

$$\begin{aligned}
p(x=0, y=0) &= (1-\theta_1)\cdot\theta_2, \\
p(x=0, y=1) &= (1-\theta_1)\cdot(1-\theta_2), \\
p(x=1, y=0) &= \theta_1\cdot(1-\theta_2), \\
p(x=1, y=1) &= \theta_1\cdot\theta_2.
\end{aligned}$$

This can be compactly written as:

$$p(x, y|\theta_1, \theta_2) = \theta_1^x(1-\theta_1)^{(1-x)}\theta_2^{x\odot y}(1-\theta_2)^{(1-x\odot y)},$$

where $\odot$ is the Exclusive NOR operator.

For question (b), the MLE for θ_1 is $\frac{4}{7}$, while that for θ_2 is $\frac{4}{7}$. Since we assumed the independency between θ_1 and θ_2, both MLE can be arrived at by simply counting. The evidence is given by:

$$p(\mathcal{D}|\theta_{\mathrm{MLE}}) = \left(\frac{4}{7}\right)^4\cdot\left(\frac{3}{7}\right)^3\cdot\left(\frac{3}{7}\right)^3\cdot\left(\frac{4}{7}\right)^4.$$

For question (c), the MLE for θ is computed by normalizing the counting vector: $(2, 1, 2, 2)$, so:

$$\theta_{\mathrm{MLE}} = \left(\frac{2}{7}, \frac{2}{7}, \frac{1}{7}, \frac{2}{7}\right).$$

The evidence is:

$$\left(\frac{2}{7}\right)^2 \cdot \left(\frac{2}{7}\right)^2 \cdot \left(\frac{1}{7}\right) \cdot \left(\frac{2}{7}\right)^2 .$$

For question (d):

```
import math
x=[1,1,0,1,1,0,0]
y=[1,0,0,0,1,0,1]
l2=0
l4=0
e=1/10**5
for shadow in range(7):
    temp1=0
    temp2=0
    temp00=0
    temp10=0
    temp01=0
    temp11=0
    for i in range(len(x)):
        if i==shadow:
            continue
        if x[i]==1:
            temp1=temp1+1
        if x[i]==y[i]:
            temp2=temp2+1
        if x[i]==0 and y[i]==0:
            temp00=temp00+1
        if x[i]==0 and y[i]==1:
            temp01=temp01+1
        if x[i]==1 and y[i]==0:
            temp10=temp10+1
        if x[i]==1 and y[i]==1:
            temp11=temp11+1
    theta_1=temp1/(len(x)-1)
    theta_2=temp2/(len(x)-1)
    s=temp00+temp01+temp10+temp11
    theta_00=temp00/s
    theta_01=temp01/s
    theta_10=temp10/s
```

```
    theta_11=temp11/s
    p2=theta_1**(x[shadow])*(1-theta_1)**(1-x[shadow])*theta_2
        **(1-x[shadow]^y[shadow])*(1-theta_2)**(x[shadow]^y[
        shadow])
    p4=theta_00**(x[shadow]==0 and y[shadow==0])*theta_01**(x[
        shadow]==0 and y[shadow==1])*theta_10**(x[shadow]==1
        and y[shadow==0])*theta_11**(x[shadow]==1 and y[shadow
        ==1])
    l2=l2+math.log(p2+e)
    l4=l4+math.log(p4+e)
print(l2)
print(l4)
```

The result is:

```
-12.136441189337646
-28.04302596169576
```

Hence the CV will pick M_2. The reason behind is that M_4 assumes zero probability for $(0, 1)$ during the cross-validation, which significantly declines the confidence.

For question (e), the BICs for M_2 and M_4 are respectively:

$$\text{BIC}(M_2, \mathcal{D}) = -11.51,$$

$$\text{BIC}(M_4, \mathcal{D}) = -12.38.$$

Hence the BIC prefers M_2 as well.

5.9 Posterior median is optimal estimate under L1 loss

The posterior loss expectation is (where we have omitted $\mathcal{D}$ w.l.o.g.):

$$\begin{aligned}\rho(a) &= \int |y-a|p(y)\mathrm{d}y = \int_{-\infty}^{a}(a-y)p(y)\mathrm{d}y + \int_{a}^{+\infty}(y-a)p(y)\mathrm{d}y \\ &= a\left\{\int_{-\infty}^{a}p(y)\mathrm{d}y - \int_{a}^{+\infty}p(y)\mathrm{d}y\right\} - \int_{-\infty}^{a}yp(y)\mathrm{d}y + \int_{a}^{+\infty}yp(y)\mathrm{d}y.\end{aligned}$$

Differentiating the loss w.r.t. a ends up with:

$$\frac{\partial}{\partial a}\rho(a) = \left\{\int_{-\infty}^{a}p(y)dy - \int_{a}^{+\infty}p(y)\right\} + a\cdot 2p(a) - a\cdot 2p(a).$$

Set it to zero so:

$$\int_{-\infty}^{a} p(y)dy = \int_{a}^{+\infty} p(y),$$

whose summation is unity, hence a satisfies:

$$\int_{-\infty}^{a} p(y)dy = \int_{a}^{+\infty} p(y) = \frac{1}{2},$$

which identifies the median.

5.10 Decision rule for trading off FPs and FNs

This is tantamount to exercise 5.2. Replacing λ_{01} by $c \cdot \lambda_{10}$, the rest follows the corresponding results.

6 Frequentist statistics

For one who has delved into Bayesian statistics before acknowledging the frequentist statistics, the latter is somewhat awkward and intimidating. However, in most scenarios regarding science and engineering, orthodox statistics remains the optimal candidate. Since when the size of the dataset grows, the privilege of using Bayes is easily overwhelmed by the asymptotic performance of classical statistics. Moreover, although Bayesians claim to have eliminated the overfitting problem with prior distributions, the subjectiveness in selecting hyperparameters also increases. This problem cannot be solved with empirical Bayesian or more complex graphical structures, since none of them cut down the degree of freedom for parameters.

For the reasons above, the frequentist statistics should receive equal attention as, if not more than, the Bayesian version.

6.1 Pessimism of LOOCV

Under the setting in this problem, the best classification method can achieve is the accuracy of 50%, since the features are independent of the labels. This is, however, the idealistic situation where we assume that

$$N \to \infty,$$

where $N = N_1 = N_2$. When N is finite, the classification accuracy usually fluctuates around 50%. For the label to be *randomly* assigned, we should have that for an arbitrary (probabilistic) algorithm/machine that terminates within a time complexity a polynomial of N_i, its classification accuracy should not deviate from 50% larger than $\frac{1}{p(N)}$, in which p is an arbitrary polynomial function. The machine in this case is referred to as a probabilistic polynomial time (PPT) machine and is of remarkable interest to cryptologists. To generate cheap pseudo-random numbers efficiently is one of the most important tasks for cryptography.

For LOOCV, if a sample for class one is omitted, then the classification probability is:

$$(\frac{N-1}{2N-1}, \frac{N}{2N-1}).$$

Hence the probability that the omitted sample is correctly labelled is:

$$\frac{N-1}{2N-1} \leq \frac{1}{2},$$

justifies its pessimism in this setting.

The dependency between features and labels paves the way for decoupling features or security/forensics in machine learning. However, it is impossible to justify that some dataset *cannot be learned* by any machines, whose converse side is extremely simple.

6.2 James Stein estimator for Gaussian means

The prior for θ_i is:

$$\mathcal{N}(\theta_i|m_0, \tau_0^2),$$

and the likelihood is given by:

$$\mathcal{N}(y_i|\theta_i, \sigma^2).$$

For question (a), we begin by integrating out θ_i and establishing the dependency of $\mathcal{D} = \{y_i\}_{i=1}^6$ on m_0 and τ_0^2:

$$\begin{aligned}
p(y|m_0, \tau_0^2) &= \int p(y, \theta|m_0, \tau_0^2)\mathrm{d}\theta \\
&= \int p(y|\theta\sigma^2) \cdot p(\theta|m_0, \tau_0^2)\mathrm{d}\theta \\
&= \frac{1}{2\pi\sqrt{\sigma^2\tau_0^2}} \cdot \int \exp\left\{-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta - m_0)^2}{\tau_0^2}\right)\right\}\mathrm{d}\theta \\
&\propto \exp\left\{-\frac{1}{2}\left(-\frac{(\frac{y}{\sigma^2} + \frac{m_0}{\tau_0^2})^2}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}} + \frac{y^2}{\sigma^2}\right)\right\} \int \exp\left\{(a \cdot \theta - b(y))^2\right\}\mathrm{d}\theta \\
&\propto \exp\left\{-\frac{1}{2}\left(-\frac{(y + \frac{m_0\sigma^2}{\tau_0^2})^2}{\sigma^2 + \frac{\sigma^4}{\tau_0^2}} + \frac{y^2}{\sigma^2}\right)\right\} \\
&= \mathcal{N}\left(y|m_0, \sigma^2 + \tau_0^2\right),
\end{aligned}$$

where we have canceled terms independent from y and completing the square in the final step of deduction. Given $\sigma^2 = 500$, we have:

$$\hat{m_0} = \bar{x} = 1527.5,$$

$$\hat{\tau_0^2} = 1878.58 - \sigma^2 = 1378.58,$$

For question (b), the posterior distribution of θ_i given y_i and other hyperparameters is:

$$\begin{aligned} p(\theta|y) &\propto p(\theta) \cdot p(y|\theta) \\ &\propto \exp\left\{-\frac{(\theta - m_0)^2}{2\tau_0^2} - \frac{(\theta - y)^2}{2\sigma^2}\right\} \\ &= \mathcal{N}(\theta|\frac{m_0\sigma^2 + y\tau_0^2}{\sigma^2 + \tau_0^2}, \frac{\sigma^2\tau_0^2}{\sigma^2 + \tau_0^2}). \end{aligned}$$

For question (c), the interval is $(\mu' - 1.96\sigma', \mu' + 1.96\sigma')$, where μ' and σ' are from question (b).

For question (d), a smaller σ^2 would reduce the ML-II into the ordinary posterior analysis. The parameter σ^2 can be understood as the *noise* on the observations. The less noise we assumed, the more precise the observations are, and the intermedia θ becomes less necessary.

6.3 $\hat{\sigma}^2_{\mathbf{MLE}}$ is biased

Consider:

$$\mathbb{E}[(x_n - \bar{x})^2] = \mathbb{E}[x_n^2] + \frac{1}{N^2}\sum_{j,k=1}^{N}\mathbb{E}[x_j x_k] - \frac{2}{N}\sum_{j=1}^{N}\mathbb{E}[x_n x_j].$$

The second term on the r.h.s. contains N components that equal to $\mathbb{E}[x^2] = \mu^2 + \sigma^2$ and $N^2 - N$ components that equal to $\mathbb{E}[x_l x_m] = \mu^2$, where $l \neq m$. For the third term, the numbers are 1 and $N - 1$, thus:

$$\mathbb{E}[(x_n - \bar{x})^2] = \frac{N-1}{N}\sigma^2,$$

so is $\hat{\sigma}^2_{\mathrm{MLE}}$. The unbiased estimation, of which one who has taken several elementary courses on probability and statistics should have been very familiar, is:

$$\frac{N}{N-1} \cdot \hat{\sigma}^2_{\mathrm{MLE}}.$$

6.4 Estimation of σ^2 when μ is known

The likelihood is:

$$p(x|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Taking logarithm and gradient:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2} \ln p(\mathcal{D}|\sigma^2) = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2.$$

Setting it to zero yields:

$$\sigma^2_{\mathrm{MLE}} = \frac{\sum_{n=1}^{N} (x_n - \mu)^2}{N}.$$

In this case the MLE is unbiased since:

$$\mathbb{E}(x_n - \mu^2) = \mathbb{E}(x_n^2) + \mu^2 - 2\mu\mathbb{E}(x_n) = \sigma^2.$$

7 Linear regression

7.1 Behavior of training set error with increasing sample size

The statement **the error on the test will always decrease as we get more training data since the model will be better estimated** is not a gold standard. The prerequisites behind this statement are at least:

- The test data shares an identical distribution with the training data. Which, although plausible in general, fails in cases such as adversarial training (where an adversary schemes for samples that hinders the model) or defense against zero-day attack (where the model is required to detect an unknown threat from benign samples solely).
- The model is complex enough to learn the knowledge embedded in the data. For a basic example, a tabular predictor that only saves the latest finite samples and predicts by table looking-up. In this case, the size of the dataset would hardly help. Moreover, the break of the i.i.d. assumption could worsen its performance.

We now reduce the discussion in cases where the assumptions for PAC learning holds. When the training set is small, the trained model is usually over-fitted to the current data set (since the model is very complex), so the accuracy can be relatively high. As the training set increases, the model has to learn to adapt to more general-purpose parameters, thus reducing the overfitting effect laterally, resulting in lower accuracy. As pointed out in Section 7.5.4, increasing the training set is an important method of countering over-fitting besides adding regularizers. Increasing dataset is fundamentally the only solution to models' robustness since one cannot exhaust all possible data augmentation strategies.

7.2 Multi-output linear regression

For multi-output linear regression, if the outputs are independent, then for each output dimension subscripted by j, we have:

$$p(y_j|\mathbf{x}, \mathbf{w}_j) = \mathcal{N}(y_j|\mathbf{x}^{\mathrm{T}}\mathbf{w}_j, \sigma_j^2),$$

then:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{W}) = \prod_{j=1}^{M} \mathcal{N}(y_j|\mathbf{x}^{\mathrm{T}}\mathbf{w}_j, \sigma_j^2).$$

The independence implies:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{n=1}^{N}\prod_{j=1}^{M} \mathcal{N}(y_{n,j}|\mathbf{x}_n^{\mathrm{T}}\mathbf{w}_j, \sigma_j^2).$$

Taking its logarithm and saving terms dependent on $\mathbf{W}$, we have:

$$\mathcal{L}(\mathbf{W}) = \sum_{n=1}^{N}\sum_{j=1}^{M} \frac{1}{\sigma_j^2}\left(y_{n,j} - \mathbf{x}_n^{\mathrm{T}}\mathbf{w}_j\right)^2.$$

Interchanging the order of summation helps to decompose the loss into:

$$\sum_{j=1}^{M} \mathcal{L}_j(\mathbf{w}_j) = \sum_{j=1}^{M} \frac{1}{\sigma_j^2}\sum_{n=1}^{N}\left(y_{n,j} - \mathbf{x}_n^{\mathrm{T}}\mathbf{w}_j\right)^2,$$

where each:

$$\mathcal{L}_j(\mathbf{w}_j) = \frac{1}{\sigma_j^2}\sum_{n=1}^{N}\left(y_{n,j} - \mathbf{x}_n^{\mathrm{T}}\mathbf{w}_j\right)^2,$$

is but the MLE loss for a 1D linear regression. Thus the columns of $\mathbf{W}$ can be estimated independently by:

$$\mathbf{w}_j^{\mathrm{MLE}} = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}_j,$$

where $\mathbf{X}$ is the $D*N$ design matrix and $\mathbf{Y}$ is a columm matrix with length N that embeds the j-th component of the output. This is a little different from the symbols from the textbook but simple α-reductions would eliminate such difference. Equation (7.90) is incorrect by missing one design matrix.

As a compact way of writing $\mathbf{W}^{\mathrm{MLE}}$, we would have:

$$\mathbf{W}^{\mathrm{MLE}} = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y},$$

where $\mathbf{Y}$ is a $N * M$ matrix.

For the case in this exercise, we have $D = 2$, $N = 6$, $M = 2$:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix},$$

$$\mathbf{Y} = \begin{pmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

Thus:

$$\mathbf{W} = \begin{pmatrix} -\frac{4}{3} & -\frac{4}{3} \\ \frac{4}{3} & \frac{4}{3} \end{pmatrix}.$$

One can observe that two columns for $\mathbf{W}$ are identical, which is obvious by examing the two columns from $\mathbf{Y}$.

7.3 Centering and ridge regression

The loss is given by:

$$\mathcal{L}(\mathbf{w}, w_0) = \left(\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{w} - w_0\mathbf{1}\right)^{\mathrm{T}} \left(\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{w} - w_0\mathbf{1}\right).$$

Taking partial gradient w.r.t. w_0 yields:

$$\frac{\partial}{\partial w_0}\mathcal{L} = \frac{\mathbf{1}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{w})}{N}.$$

Setting it to zero gives (7.94), where we have made use of $\bar{x} = 0$.

Taking partial gradient w.r.t. $\mathbf{w}$ yields:

$$\mathbf{w}^{\mathrm{MLE}} = (\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}\mathbf{X}(\mathbf{y} - w_0\mathbf{1}).$$

In the equation (7.95), the observed data has to be centralized.

7.4 MLE for σ^2 for linear regression

The likelihood is given by:

$$\begin{aligned} p(\mathcal{D}|\mathbf{w}, \sigma^2) =& p(\mathbf{Y}|\mathbf{w}, \sigma^2, \mathbf{X}) \\ =& \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) \\ =& \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{w}^{\mathrm{T}}\mathbf{x}_n, \sigma^2) \\ =& \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2\right\}. \end{aligned}$$

As for σ^2:

$$\frac{\partial}{\partial\sigma^2}\log p(\mathcal{D}|\mathbf{w}, \sigma^2) = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(y_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2.$$

Setting it to zero yields:

$$\sigma^2_{\mathrm{MLE}} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2.$$

7.5 MLE for the offset term in linear regression

For (7.97), we only have to use the second equation derived in exercise 7.3.

For (7.98), recall the third equation that we have derived in exercise 7.3.

7.6 MLE for simple linear regression

For (7.99), we plug the $D = 1$ assumption into (7.98) so:

$$w_1^{\mathrm{MLE}} = \frac{\sum_{n=1}^{N}(y_n - \bar{y})(x_n - \bar{x})}{\sum_{n=1}^{N}(x_n - \bar{x})^2}.$$

Recall that $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and that for $\bar{y}$, so (7.99) is proven.

For (7.100), use (7.97) directly.

7.7 Sufficient statistics for online linear regression

For question (a), according to (7.99), w_1 can be estimated from $C_{xy}^{(n)}$ and $C_{xx}^{(n)}$ solely.

For question (b), according to (7.100), w_0 can be estimated from $\bar{x}^{(n)}$, $\bar{y}^{(n)}$ and w_1. Hence $\bar{x}^{(n)}$, $\bar{y}^{(n)}$, $C_{xy}^{(n)}$ and $C_{xx}^{(n)}$ are necessary.

For question (c), the solution has been given by (7.103)-(7.104). For y, the matter is simply an α-reduction.

For question (d), we need to prove:

$$(n+1)C_{xy}^{(n+1)} = nC_{xy}^{(n)} + x_{n+1}y_{n+1} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}.$$

Expand the C_{xy} in both two sides then the l.h.s. becomes:

$$\sum_{i=1}^{n+1}(x_i - \bar{x}^{(n+1)})(y_i - \bar{y}^{(n+1)}),$$

the r.h.s. becomes:

$$\sum_{i=1}^{n}(x_i - \bar{x}^{(n)})(y_i - \bar{y}^{(n)}) + x_{n+1}y_{n+1} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}.$$

Finally, substituting the average at the $(n+1)$-th iteration in the l.h.s. and r.h.s. by (7.104), this is sufficient for arriving in (7.105).

For question (e) and (f):

```
import math
import matplotlib.pyplot as plt
x=[94,96,94,95,104,106,108,113,115,121,131]
y=[0.47,0.75,0.83,0.98,1.18,1.29,1.40,1.60,1.75,1.90,2.23]
bx=(x[0]+x[1])/2
by=(y[0]+y[1])/2
Cxx=((x[0]-bx)**2+(x[1]-bx)**2)/2
Cxy=((y[0]-by)*(x[0]-bx)+(y[1]-by)*(x[1]-bx))/2
w1=[]
w0=[]
for n in list(range(2,11)):
    bx_=bx+(x[n]-bx)/(n+1)
    by_=by+(y[n]-by)/(n+1)
    Cxx=(x[n]*x[n]+n*Cxx+n*bx*bx-(n+1)*bx_*bx_)/(n+1)
```

```
    Cxy=(x[n]*y[n]+n*Cxy+n*bx*by-(n+1)*bx_*by_)/(n+1)
    bx=bx_
    by=by_
    w1.append(Cxy/Cxx)
    w0.append(by-w1[n-2]*bx)
```

With: Where we borrow data from exercise 7.8 for illustration. The data

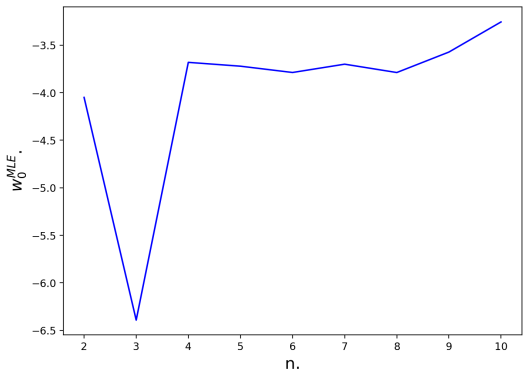


Figure. 4. Exercise 7.7. P1

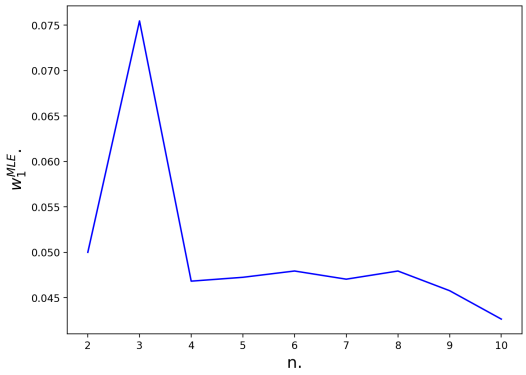


Figure. 5. Exercise 7.7. P2

in x and y appear in sequential order for Figure. 7. 14. That is to say, if we shuffle x and y accordingly, the estimation for weights is expected to converge faster.

7.8 Bayesian linear regression in 1d with known σ^2

For question (a), the estimation for σ^2 is 0.3173.

For question (b), we have:

$$p(\mathbf{w}) \propto \mathcal{N}(w_1|0,1) \propto \exp\left\{-\frac{1}{2}w_1^2\right\}.$$

To simplify the algebra, we observe that w_1 and w_0 are independent in this prior. Thus the prior distribution can be reduced to: $\mathcal{N}(w_1|0,1) \cdot \mathcal{N}(w_0|\gamma,\infty)$, where γ can be an arbitrary finite number, thus:

$$\mathbf{m}_0 = \begin{pmatrix} 0 \\ \gamma \end{pmatrix},$$

$$\mathbf{V}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

For question (c) and (d), we consider the posterior distribution for parameters:

$$\begin{aligned} p(\mathbf{w}|\mathcal{D},\sigma^2) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0,\mathbf{V}_0) \cdot \prod_{n=1}^{N} \mathcal{N}(y_n|w_0 + w_1 x_n, \sigma^2) \\ &\propto \exp\left\{-\frac{w_1^2}{2}\right\} \cdot \prod_{n=1}^{N} \exp\left\{-\frac{(y_n - w_0 - w_1 x_n)^2}{2\sigma^2}\right\}, \end{aligned}$$

where we only maintain terms dependent on w_0 and w_1. To marginalize out w_0, we have:

$$\begin{aligned} p(w_1|\mathcal{D},\sigma^2) &= \int p(w_1, w_0) \mathrm{d}w_0 \\ &\propto \exp\left\{-\frac{w_1^2}{2}\right\} \cdot \int \exp\left\{Aw_1^2 + Bw_0^2 + Cw_0w_1 + Dw_1 + Ew_0 + F\right\} \mathrm{d}w_0 \\ &= \exp\left\{-\frac{w_1^2}{2} + Aw_1^2 + Dw_1 + F\right\} \cdot \int \exp\left\{Bw_0^2 + Cw_0w_1 + Ew_0\right\} \mathrm{d}w_0 \\ &= \exp\left\{-\frac{w_1^2}{2} + Aw_1^2 + Dw_1 + F - \frac{(Cw_1 + E)^2}{4B}\right\} \cdot \\ &\int \exp\left\{Bw_0^2 + (Cw_1 + E)w_0 + \frac{(Cw_1 + E)^2}{4B}\right\} \mathrm{d}w_0 \\ &\propto \exp\left\{-\frac{w_1^2}{2} + Aw_1^2 + Dw_1 + F - \frac{(Cw_1 + E)^2}{4B}\right\}. \end{aligned}$$

Hence the posterior distribution over w_1 is a normal distribution. The coefficients for w_1^2 and w_1 in the exponential are respectively:

$$-\frac{1}{2} + A - \frac{C^2}{4B},$$
$$D - \frac{CE}{2B}.$$

Thence its posterior variance is:

$$\frac{1}{1 - 2A + \frac{C^2}{2B}},$$

its mean is:

$$\frac{2BD - CE}{2B - 4AB + C^2}.$$

Finally, let us plug A to E with statistics of $\mathcal{D}$:

$$A = -\frac{\sum_{n=1}^{N} x_n^2}{2\sigma^2},$$
$$B = -\frac{N}{2\sigma^2},$$
$$C = -\frac{\sum_{n=1}^{N} x_n}{\sigma^2},$$
$$D = \frac{\sum_{n=1}^{N} x_n y_n}{\sigma^2},$$
$$E = \frac{\sum_{n=1}^{N} y_n}{\sigma^2}.$$

The posterior variance is:

$$\frac{\sigma^2}{\sigma^2 + \sum x^2 - \frac{(\sum x)^2}{N}},$$

from which we observe that, with N grows, the denominator increases as the bound from the Cauchy inequality:

$$\sum_{n=1}^{N} x_n^2 - \frac{(\sum_{n=1}^{N} x_n)^2}{N} \geq 0.$$

To put in other words, The larger the Cauchy difference is, the more confidence we have for the estimation of w_1. Such difference is determined by the variance of the distribution on x. The posterior variance can be reduced to:

$$\frac{\sigma^2}{\sigma^2 + N\mathrm{var}(x)}.$$

Therefore for any fixed generative distribution on x, the uncertainty on w_1 declines as $\mathcal{O}\left\{\frac{1}{N}\right\}$.

7.9 Generative model for linear regression

For question (a), assume that $\mathbf{x}$ and y are jointly subject to a Gaussian:

$$\mathcal{N}\left(\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix} | \begin{pmatrix}\mu_{\mathbf{x}}\\ \mu_y\end{pmatrix}, \begin{pmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{xy}^{\mathrm{T}} & \Sigma_{yy}\end{pmatrix}\right).$$

We know from (4.68) that the marginal distribution of $\mathbf{x}$ and y are $\mathcal{N}(\mathbf{x}|\mu_{\mathbf{x}}, \Sigma_{xx})$ and $\mathcal{N}(\mathbf{y}|\mu_y, \Sigma_{yy})$ respectively, this is sufficient for MLE $\mu_{\mathbf{x}}$, μ_y, Σ_{xx} and Σ_{yy}:

$$\begin{aligned}
\mu_{\mathbf{x}} &= \frac{1}{N} \cdot \sum_{n=1}^{N} \mathbf{x}_n,\\
\mu_y &= \frac{1}{N} \cdot \sum_{n=1}^{N} y_n,\\
\Sigma_{xx} &= \frac{1}{N} \cdot \sum_{n=1}^{N} (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^{\mathrm{T}} = \frac{1}{N} \cdot \bar{\mathbf{X}}\bar{\mathbf{X}}^{\mathrm{T}},\\
\Sigma_{yy} &= \frac{1}{N} \cdot \sum_{n=1}^{N} (y_n - \bar{y})^2 = \frac{1}{N}\bar{\mathbf{Y}}^{\mathrm{T}}\bar{\mathbf{Y}}.
\end{aligned}$$

For an estimation of Σ_{xy}, connecting $\mathbf{x}$ and y together then estimating this vector's covariance matrix (this is an ordinary Gaussian model), picking only the first D components from its last column:

$$\Sigma_{xy} = \frac{1}{N}\bar{\mathbf{X}}\bar{\mathbf{Y}}.$$

Finally, plugging our observations into (4.69):

$$\mu_{y|\mathbf{x}} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\mathbf{x} - \mu_{\mathbf{x}}),$$

which equals:

$$\bar{y} + \frac{1}{N} \cdot \bar{\mathbf{Y}}^{\mathrm{T}}\bar{\mathbf{X}}^{\mathrm{T}} \cdot N \cdot (\bar{\mathbf{X}}\bar{\mathbf{X}}^{\mathrm{T}})^{-1} \cdot (\mathbf{x} - \bar{\mathbf{x}}).$$

This is identical to what (7.109)-(7.111) implies, hence the proof is completed.

For question (b), there is no significant difference except that we are now equipped with a distribution of x. This enables us the ability of generating more samples and sheds light on active querying strategies, in which we can query an oracle for samples in order to accelerate the convergence of the learning process.

7.10 Bayesian linear regression using the g-prior

For Bayesian linear regression model, the likelihood is as always:

$$p(\mathcal{D}|\mathbf{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{w}^{\mathrm{T}}\mathbf{x}_n,\sigma^2).$$

The prior distribution is Gaussian-Inverse Gamma distribution:

$$\begin{aligned}
p(\mathbf{w},\sigma^2) &= \mathrm{NIG}(\mathbf{w},\sigma^2|\mathbf{w}_0,\mathbf{V}_0,a_0,b_0) \\
&= \mathcal{N}(\mathbf{w}|\mathbf{w}_0,\sigma^2\mathbf{V}_0)\cdot \mathrm{IG}(\sigma^2|a_0,b_0) \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\sigma^2\mathbf{V}_0|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T(\sigma^2\mathbf{V}_0)^{-1}(\mathbf{w}-\mathbf{w}_0)\right\}\cdot \\
&\quad \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)}\exp\left\{-\frac{b_0}{\sigma^2}\right\} \\
&= \frac{b_0^{a_0}}{(2\pi)^{\frac{D}{2}}|\mathbf{V}_0|^{\frac{1}{2}}\Gamma(a_0)}(\sigma^2)^{-\left(a_0+\frac{D}{2}+1\right)}\cdot\exp\left\{-\frac{(\mathbf{w}-\mathbf{w}_0)^{\mathrm{T}}\mathbf{V}_0^{-1}(\mathbf{w}-\mathbf{w}_0)+2b_0}{2\sigma^2}\right\}.
\end{aligned}$$

The posterior distribution takes the form:

$$\begin{aligned}
p(\mathbf{w},\sigma^2|\mathcal{D}) &\propto p(\mathbf{w},\sigma^2)p(\mathcal{D}|\mathbf{w},\sigma^2) \\
&\propto \frac{b_0^{a_0}}{(2\pi)^{\frac{D}{2}}|\mathbf{V}_0|^{\frac{1}{2}}\Gamma(a_0)}(\sigma^2)^{-\left(a_0+\frac{D}{2}+1\right)}\cdot \\
&\quad \exp\left\{-\frac{(\mathbf{w}-\mathbf{w}_0)^{\mathrm{T}}\mathbf{V}_0^{-1}(\mathbf{w}-\mathbf{w}_0)+2b_0}{2\sigma^2}\right\}\cdot \\
&\quad (\sigma^2)^{-\frac{N}{2}}\cdot\exp\left\{-\frac{\sum_{n=1}^{N}(y_n-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2}{2\sigma^2}\right\}.
\end{aligned}$$

To decompose the updated hyperparameters from this form, we have to find:

- The exponential of σ^2.
- The squared term within the exponential.

The exponential of σ^2 in the posterior is:

$$-\left(a_0+\frac{D}{2}+1\right)-\frac{N}{2},$$

thus we have:

$$a_N=a_0+\frac{N}{2}=\frac{N}{2}.$$

The coefficient of $\mathbf{w}^{\mathrm{T}}\mathbf{w}$ in the exponential (as the introducer matrix of the inner product) is:

$$-\frac{\mathbf{V}_0^{-1}}{2\sigma^2}-\frac{\mathbf{X}\mathbf{X}^{\mathrm{T}}}{2\sigma^2},$$

therfore:

$$\mathbf{V}_N^{-1}=\mathbf{V}_0^{-1}+\mathbf{X}\mathbf{X}^{\mathrm{T}},$$

with $\mathbf{V}_0=g(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}$:

$$\mathbf{V}_N=\frac{g}{g+1}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}.$$

The coefficient of $\mathbf{w}^{\mathrm{T}}$ in the exponential is:

$$\frac{\mathbf{V}_0^{-1}\mathbf{w}_0}{\sigma^2}+\frac{\sum_{n=1}^{N}y_n\mathbf{x}_n}{\sigma^2}=\frac{\mathbf{V}_0^{-1}\mathbf{w}_0}{\sigma^2}+\frac{\mathbf{X}\mathbf{Y}}{\sigma^2}.$$

So we have:

$$\mathbf{V}_0^{-1}\mathbf{w}_0+\mathbf{X}\mathbf{Y}=\mathbf{V}_N^{-1}\mathbf{w}_N.$$

This yields:

$$\begin{aligned}\mathbf{w}_N&=\mathbf{V}_N\left(\mathbf{V}_0^{-1}\mathbf{w}_0+\mathbf{X}\mathbf{Y}\right)\\&=\frac{g}{g+1}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\left(\frac{\mathbf{X}\mathbf{X}^{\mathrm{T}}}{g}\mathbf{0}+\mathbf{X}\mathbf{Y}\right)\\&=\frac{g}{g+1}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{X}\mathbf{Y}.\end{aligned}$$

Finally, completing the square within the exponential yields:

$$(\mathbf{w}-\mathbf{w}_0)^{\mathrm{T}}\mathbf{V}_0^{-1}(\mathbf{w}-\mathbf{w}_0)+2b_0+(\mathbf{w}^{\mathrm{T}}\mathbf{X}-\mathbf{Y})^2=(\mathbf{w}-\mathbf{w}_N)^{\mathrm{T}}\mathbf{V}_N^{-1}(\mathbf{w}-\mathbf{w}_N)+2b_N.$$

Plugging in what we have already known about $\mathbf{w}_N$ and $\mathbf{V}_N$ results in:

$$b_N=\frac{\mathbf{Y}^{\mathrm{T}}\mathbf{Y}}{2}+\frac{g}{2(g+1)}\mathbf{w}_{\mathrm{MLE}}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{w}_{\mathrm{MLE}}.$$

Now we have (7.113)-(7.116) established.

8 Logistic regression

Logistic regression is the basic form of classification, which is the fundamental task of machine intelligence. Even for classifying complex data structures, the last layers of the model are usually logistic structures.

8.1 Spam classification using logistic regression

Practice by yourself. I know nothing about MATLAB.

8.2 Spam classification using naive Bayes

Practice by yourself.

8.3 Gradient and Hessian of log-likelihood for logistic regression

For question (a),

$$\frac{\mathrm{d}}{\mathrm{d}a}\sigma(a) = \frac{\exp(-a)}{(1+\exp(-a))^2} = \frac{1}{1+\mathrm{e}^{-a}}\frac{\mathrm{e}^{-a}}{1+\mathrm{e}^{-a}} = \sigma(a)\cdot(1-\sigma(a)).$$

For question (b),

$$\begin{aligned} g(\mathbf{w}) =& \frac{\partial}{\partial \mathbf{w}}\mathrm{NLL}(\mathbf{w}) \\ =& -\sum_{n=1}^{N}\frac{\partial}{\partial \mathbf{w}}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)] \\ =& -\sum_{n=1}^{N} y_i \frac{1}{\sigma_i}\sigma_i(1-\sigma_i)\cdot\mathbf{x}_i + (1-y_i)\frac{-1}{1-\sigma_i}\sigma(1-\sigma_i)\cdot\mathbf{x}_i \\ =& \sum_{n=1}^{N}(\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i) - y_i)\mathbf{x}_i, \end{aligned}$$

where $\sigma_i = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)$.

For question (c), the result is obvious. For an arbitrary vector $\mathbf{u}$:

$$\begin{aligned}\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} &= (\mathbf{X}\mathbf{u})^{\mathrm{T}}\,\mathbf{S}\,(\mathbf{X}\mathbf{u}) \\ &= \mathbf{v}^{\mathrm{T}}\mathbf{S}\mathbf{v} \\ &= \sum_{d=1}^{D} v_d^2 \cdot \mu_d \cdot (1-\mu_i) \geq 0.\end{aligned}$$

Hence $\mathbf{H}$ is positive definite if all μ_i are within $(0,1)$, otherwise it is semi-positive definite.

8.4 Gradient and Hessian of log-likelihood for multinomial logistic regression

For question (a), given a sample indexed i we have,

$$\mu_{ik} = \frac{\exp(\mathbf{w}_k^{\mathrm{T}}\mathbf{x}_i)}{\sum_c \exp(\mathbf{w}_c^{\mathrm{T}}\mathbf{x}_i)},$$

$$\eta_{ij} = \mathbf{w}_j^{\mathrm{T}}\mathbf{x}_i.$$

Now we have:

$$\begin{aligned}\frac{\partial \mu_{ik}}{\partial \eta_{ij}} &= \frac{\frac{\partial \exp(\eta_{ik})}{\partial \eta_{ij}} \cdot \sum_c \exp(\eta_{ic}) - \frac{\partial \sum_c \exp(\eta_{ic})}{\partial \eta_{ij}} \cdot \exp(\eta_{ik})}{\left(\sum_c \exp(\eta_{ic})\right)^2} \\ &= \frac{\exp(\eta_{ik}) \cdot \delta_{kj} \cdot \sum_c \exp(\eta_{ic}) - \exp(\eta_{ij}) \cdot \exp(\eta_{ik})}{\left(\sum_c \exp(\eta_{ic})\right)^2} \\ &= \mu_{ik} \cdot \delta_{kj} - \mu_{ij} \cdot \mu_{ik},\end{aligned}$$

what dominates is but the elementary calculus.

For question (b), recall that:

$$l(\mathbf{W}) = \sum_{i=1}^{N} \sum_c y_{ic} \cdot \log \mu_{ic}.$$

Let $l_i(\mathbf{W}) = \sum_c y_{ic} \cdot \log \mu_{ic}$, we are now ready for reduction:

$$\begin{aligned}
\frac{\partial l_i}{\partial \mathbf{w}_j} &= \frac{\partial}{\partial \mathbf{w}_j} \sum_c y_{ic} \cdot \log \mu_{ic} \\
&= \sum_c \frac{y_{ic}}{\mu_{ic}} \cdot \frac{\partial \mu_{ic}}{\partial \eta_{ij}} \cdot \frac{\partial \eta_{ij}}{\partial \mathbf{w}_j} \\
&= \sum_c \frac{y_{ic}}{\mu_{ic}} \cdot \mu_{ic} \cdot (\delta_{cj} - \mu_{ij}) \cdot \mathbf{x}_i \\
&= \sum_c y_{ic} \cdot (1 - \mu_{ij}) \cdot \mathbf{x}_i \\
&= y_{ij} \cdot (1 - \mu_{ij}) \cdot \mathbf{x}_i - \sum_{c \neq j} y_{ic} \cdot \mu_{ij} \cdot \mathbf{x}_i \\
&= y_{ij} \cdot (1 - \mu_{ij}) \cdot \mathbf{x}_i + (y_{ij} - 1) \cdot \mu_{ij} \cdot \mathbf{x}_i \\
&= (y_{ij} - \mu_{ij}) \cdot \mathbf{x}_i.
\end{aligned}$$

Summarizing over i yields (8.126).

For question (c), we have by definition:

$$\mathbf{H}_{c,c'} = \nabla_{\mathbf{w}_{c'}} \nabla_{\mathbf{w}_c} l(\mathbf{W}).$$

Hence we begin with the result from question (b):

$$\begin{aligned}
\nabla_{\mathbf{w}_{c'}} \nabla_{\mathbf{w}_c} l_i &= \frac{\partial}{\partial \mathbf{w}_{c'}} (y_{ic} - \mu_{ic}) \cdot \mathbf{x}_i \\
&= -\frac{\partial}{\partial \mathbf{w}_{c'}} \mu_{ic} \cdot \mathbf{x} \\
&= -\frac{\partial \mu_{ic}}{\partial \eta_{ic'}} \frac{\partial \eta_{ic'}}{\partial \mathbf{w}_{c'}} \cdot \mathbf{x}_i \\
&= -\mu_{ic}(\delta_{cc'} - \mu_{ic'}) \cdot \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}},
\end{aligned}$$

where in the last step we have to adopt the outer product to span the Hessian. Summarizing over i yields the desired result (8.127).

8.5 Symmetric version of l2 regularized multinomial logistic regression

We borrow the results from exercise 8.5. Once the l_2 prior is introduced, the likelihood becomes:

$$l_2 = l - \sum_c \lambda \mathbf{w}_c^{\mathrm{T}} \mathbf{w}_c.$$

Therefore (8.126) becomes:

$$\nabla_{\mathbf{w}_c} l_2 = \sum_i (y_{ic} - \mu_{ic}) \cdot \mathbf{x}_i - \lambda \mathbf{w}_c.$$

At the unique optimum we have $\forall c$,

$$\nabla_{\mathbf{w}_c} l_2 = 0,$$

which is identical to:

$$\hat{w}_{c,j} = \frac{1}{\lambda} \sum_i (y_{ic} - \mu_{ic}) \cdot x_{ij}.$$

Therefore:

$$\sum_c \hat{w}_{c,j} = \frac{1}{\lambda} \sum_i \left[\sum_c (y_{ic} - \mu_{ic}) \right] x_{ij},$$

whose value is zero since $\sum_c y_{ic} = \sum_c \mu_{ic} = 1$.

8.6 Elementary properties of l2 regularized logistic regression

For question (a), the Hessian of l_2 regularized negative log likelihood is:

$$\mathbf{H} + \lambda \mathbf{I},$$

where $\mathbf{H}$, following the derivation in exercise 8.3, is at least semi-positive definite. So the Hessian for this model is strictly (for non-trivial λ) positive definite, there is a unique optimal solution. The answer is False.

For question (b), the result is not necessarily true. For a sparse optimum, one should resort to the LASSO model, where a Laplace prior is exerted on weights.

For question (c), if $\lambda = 0$ then the model reduces to ordinary logistic regression. If the dataset is linearly separable then there exists $\mathbf{w}'$ such that $\forall i$:

$$\mathbf{w}'^{\mathrm{T}} \mathbf{x}_i \cdot y_i \geq 0.$$

Now for an arbitrary number $\alpha > 0$, the weights $\alpha \cdot \mathbf{w}'$ also meets the separation condition. Let $\alpha \to \infty$ justifies that the statement in (c) is True.

For question (d), the statement is False since the model now has to trade-off between fitness and prior knowledge. Concretely, we can prove the other statement: *as we increase λ, the likelihood of the training dataset monotonically decreases.*

Assume that $\hat{\mathbf{w}}_1$ minimize the (8.131) with λ_1, denoted by J_1. Now increase λ_1 to λ_2, we are now optimizing the loss:

$$J_2(\mathbf{w}) = -l(\mathbf{w}, \mathcal{D}_{\text{train}}) + \lambda_2 \mathbf{w}^{\mathrm{T}}\mathbf{w},$$

whose optimal solution is denoted by $\hat{\mathbf{w}}_2$.

If $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$ then we already have:

$$l(\hat{\mathbf{w}}_1, \mathcal{D}_{\text{train}}) = l(\hat{\mathbf{w}}_2, \mathcal{D}_{\text{train}}).$$

Otherwise, we would have:

$$-l(\hat{\mathbf{w}}_1, \mathcal{D}_{\text{train}}) + \lambda_2 \hat{\mathbf{w}}_1^{\mathrm{T}}\hat{\mathbf{w}}_1 > -l(\hat{\mathbf{w}}_2, \mathcal{D}_{\text{train}}) + \lambda_2 \hat{\mathbf{w}}_2^{\mathrm{T}}\hat{\mathbf{w}}_2.$$

If $l(\hat{\mathbf{w}}_2, \mathcal{D}_{\text{train}}) > l(\hat{\mathbf{w}}_1, \mathcal{D}_{\text{train}})$ then we would have:

$$\Delta = l(\hat{\mathbf{w}}_2, \mathcal{D}_{\text{train}}) - l(\hat{\mathbf{w}}_1, \mathcal{D}_{\text{train}}) > \lambda_2(\hat{\mathbf{w}}_2^{\mathrm{T}}\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1^{\mathrm{T}}\hat{\mathbf{w}}_1).$$

Finally, consider:

$$J_1(\hat{\mathbf{w}}_1) - J_1(\hat{\mathbf{w}}_2) = \Delta + \lambda_1(\hat{\mathbf{w}}_1^{\mathrm{T}}\hat{\mathbf{w}}_1 - \hat{\mathbf{w}}_2^{\mathrm{T}}\hat{\mathbf{w}}_2).$$

If $\hat{\mathbf{w}}_1^{\mathrm{T}}\hat{\mathbf{w}}_1 > \hat{\mathbf{w}}_2^{\mathrm{T}}\hat{\mathbf{w}}_2$ then $\hat{\mathbf{w}}_1$ is not the optimum of J_1 and we arrive in a contradiction. Otherwise:

$$\lambda_1(\hat{\mathbf{w}}_2^{\mathrm{T}}\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1^{\mathrm{T}}\hat{\mathbf{w}}_1) < \lambda_2(\hat{\mathbf{w}}_2^{\mathrm{T}}\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1^{\mathrm{T}}\hat{\mathbf{w}}_1) < \Delta.$$

Hence we still have the optimality of $\hat{\mathbf{w}}_1$ fail. This finishes the proof.

For question (e), the statement is False. This can be easily shown by imagine $\lambda \to \infty$.

8.7 Regularizing separate terms in 2d logistic regression

The decision boundary is given by:

$$w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 = 0,$$

hence a straight line in the $X_1 - X_2$ plane.

For question (a), the decision boundary is an arbitrary line.

For question (b), the decision boundary is a line that passes the origin.

For question (c), the decision boundary is a line parallel to the X_1 axis.

For question (c), the decision boundary is a line parallel to the X_2 axis.

9 Generalized linear models and the exponential family

9.1 Conjugate prior for univariate Gaussian in exponential family form

Recall that the 1d Gaussian distribution is:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

Rewrite it into the standard exponential family form:

$$p(x|\mu,\sigma^2) = \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \left\{\frac{\mu^2}{2\sigma^2} + \frac{\ln(2\pi\sigma^2)}{2}\right\}\right\}.$$

With $\lambda = \frac{1}{\sigma^2}$, denote:

$$\begin{aligned}\theta &= (-\frac{\lambda}{2}, \lambda\mu)^{\mathrm{T}},\\ \phi(x) &= (x^2, x)^T,\\ A(\theta) &= \frac{\lambda\mu^2}{2} + \frac{\ln(2\pi)}{2} - \frac{\ln\lambda}{2}.\end{aligned}$$

Now consider the likelihood w.r.t. a dataset $\mathcal{D} = \{x_n\}_{n=1}^N$:

$$\log p(\mathcal{D}|\theta) = \theta^{\mathrm{T}}\left(\sum_{n=1}^N \phi(x_n)\right) - N \cdot A(\theta).$$

The prior distribution of θ should satisfy the following variational form:

$$\begin{aligned}p(\theta|\mathbf{v}, M) &= \exp\left\{\theta^{\mathrm{T}}\mathbf{v} + M \cdot A(\theta)\right\}\\ &= \exp\left\{\theta_1 \cdot v_1 + \theta_2 \cdot v_2 + M \cdot A(\theta)\right\}\\ &= \exp\left\{-\frac{\lambda v_1}{2} + \lambda\mu v_2 + M \cdot \left(\frac{\lambda\mu^2}{2} + \frac{\ln(2\pi)}{2} - \frac{\ln\lambda}{2}\right)\right\}\\ &= \exp\left\{-\frac{\lambda v_1}{2} - \frac{M\ln\lambda}{2} + \lambda\mu v_2 + \frac{M\lambda\mu^2}{2}\right\} \cdot \exp\left\{\frac{\ln 2\pi}{2}\right\}\\ &\propto \exp\left\{-\frac{\lambda v_1}{2} - \frac{M\ln\lambda}{2}\right\} \cdot \exp\left\{\lambda\mu v_2 + \frac{M\lambda\mu^2}{2}\right\}.\end{aligned}$$

The first term, in which λ is the target variable, takes the form of a Gamma distribution since:

$$\mathrm{Ga}(\lambda|\alpha,\beta) \propto \exp\left\{-\beta\lambda + (\alpha-1)\ln\lambda\right\}.$$

The second term is just another Gaussian distribution since the sufficient statistics are μ and μ^2. Combine these two observations together, we have:

$$p(\theta|\mathbf{v}, M) = \text{Ga}(\lambda|\alpha, \beta) \cdot \mathcal{N}(\mu|\gamma, \tau^2).$$

The transformations between variables are:

$$\begin{aligned}
\beta &= \frac{v_1}{2}, \\
\alpha &= 1 - \frac{M}{2}, \\
\tau^2 &= -\frac{1}{M\lambda}, \\
\gamma &= -\frac{v_2}{M}.
\end{aligned}$$

In case the variance of the prior for μ is written in the information form, the precision can be written by: $\lambda \cdot (2\alpha - 2)$.

9.2 The MVN is in the exponential family

The sufficient statistics for $\mathbf{x}$ is:

$$\phi(\mathbf{x}) = (x_1, x_2, \cdots, x_M, x_1x_1, \cdots, x_Mx_M)^{\text{T}},$$

including altogether $M + M^2$ components.

Expand the information form for a Gaussian, the linear terms within the exponential are:

$$-\frac{1}{2}\sum_{m,m'} \Lambda_{m,m'}x_m x_{m'} + \sum_m x_m \sum_{m'} \Lambda_{m,m'}\mu_{m'},$$

hence the parameters can be collected by:

$$\theta = (\sum_{m'} \Lambda_{1,m'}\mu_{m'}, \cdots, \sum_{m'} \Lambda_{M,m'}\mu_{m'}, \Lambda_{1,1}, \cdots, \Lambda_{M,M}).$$

θ contains exactly the same information as μ, Σ. Finally, we have:

$$A(\theta) = \frac{1}{2}\mu^{\text{T}}\Lambda\mu + \frac{M}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma|.$$

10 Directed graphical models(Bayes nets)

Graphical models are at the central position for knowledge-based machine learning. Since with graph, most practical relationships among structured variables can be described, learned and crystallized. Although graphical models are not influential regarding more complex scenarios, it remains the most important element in this textbook.

10.1 Marginalizing a node in a DGM

Since Figure 10.14.(a) implies:

$$\begin{aligned} p(A,B,C,D,E,F,X) &= \sum_x p(A,B,C,D,E,F,x) \\ &= p(A)\cdot p(B)\cdot p(C)\cdot p(D)\sum_x p(x|A,B)\cdot p(E|C,x)\cdot p(F|x,D) \\ &= p(A)\cdot p(B)\cdot p(C)\cdot p(D)\cdot f(A,B,C,D,E,F). \end{aligned}$$

Thus the marginalized graph has to decompose in such a way with clique $\{A,B,C,D,E,F\}$. One option is as follows: The edge from E to F is neces-

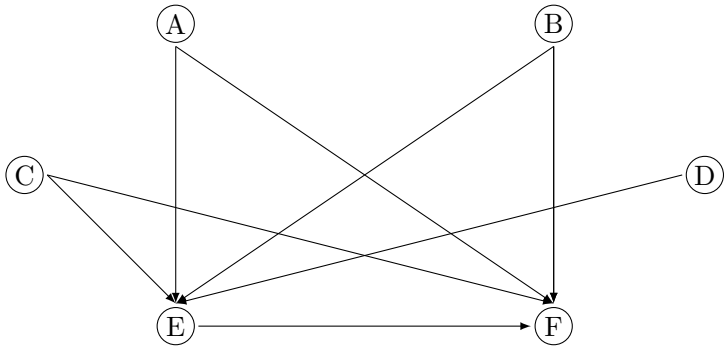


Figure. 6. Exercise 10.1.

sary since the joint probability cannot be decomposed into $f(A,B,C,D,E)\cdot g(A,B,C,D,F)$.

10.2 Bayes Ball

For question (a), for node C, there is a path:

$$C \to A,$$

so C is not independent from A given B.

For node D, there is a path:

$$D \to G \to E \to C \to A,$$

so D is not independent from A given B.

For node E, there is a path:

$$E \to C \to A,$$

so E is not independent from A given B.

For node F, there is a path:

$$F \to C \to A,$$

so F is not independent from A given B.

For node G, there is a path:

$$G \to E \to C \to A,$$

so G is not independent from A given B.

For node H, there is a path:

$$H \to F \to C \to A,$$

so H is not independent from A given B.

For node I, all pathes from I to A have to go through a collider whose intermedium is not shadowed (G and H), hence I is independent from A given B.

For question (b), for node B, there exist a path to A,

$$B \to D \to G \to A,$$

which contains a collider structure that is not blocked since G's child, J, is shadowed. Hence B is not independent from A given J.

For node C, each of its pathes to A must bypass $F \to I \to E$, which is a blocked collider, hence C is conditionally independent from A.

For node D, the result is the same as B.

For E, pathes:

$$E \to B \to D \to G \to A,$$

is unblocked, hence E is not independent from A given J.

For F, the result is the same as C.

For G, it is obviously not independent from A given J.

For H, path:

$$H \to D \to G \to A,$$

is unblocked, hence H is not independent from A given J.

For I, the path:

$$I \to E \to B \to D \to G \to A,$$

is unblocked, hence I is not independent from A given J.

10.3 Markov blanket for a DGM

The trick in the required reduction is to partition all variables into:

$$p(\mathcal{X}) = p(\mathcal{A}) \cdot p(X_i|\mathrm{Pa}(X_i)) \prod_{Y_j \in \mathrm{ch}(X_i)} p(Y_j|\mathrm{Pa}(Y_j)) \cdot p(\mathcal{B}|\mathcal{A}, \mathcal{Y}),$$

where $\mathcal{Y} = \mathrm{ch}(X_i)$, $\mathcal{A}$ contains the collection of all variable that are topological smaller than X_i, hence $\mathrm{Pa}(X_i) \subset \mathcal{A}$. Moreover, we have $\mathcal{A}$ incorporate:

$$\cup_{Y_j \in \mathrm{ch}(X_i)} \mathrm{Pa}(Y_j) - \{X_i\} - \cup_{Y_j \in \mathrm{ch}(X_i)} \mathrm{Pa}\{Y_j\}.$$

So all elements of $\mathcal{A}$ can be safely evaluated before X_i. Then all X_i's children can be computed according to a topological order. Finally, the rest variables $\mathcal{B}$ can be evaluated.

We have:

$$\begin{aligned} p(X_i|X_{-i}) &= \frac{p(\mathcal{X})}{p(X_{-i})} \\ &= \frac{p(\mathcal{X})}{\sum_{x_i} p(\mathcal{X}, X_i = x_i)}. \end{aligned}$$

Eliminating the terms $p(\mathcal{A})$ and $p(\mathcal{B}|\mathcal{A}, \mathcal{Y})$ from both the numerator and the denominator we have (10.58).

10.4 Hidden variables in DGMs

For question (a), $p(X_i)$ where $i = 1, 2, 3$ include three parameters. The term $p(H = h|X_{1:3})$ include $2^3 = 8$ free parameters. Each $p(X_i|H = h)$, where $i = 4, 5, 6$ includes two parameters. Thus we have:

$$3 + 8 + 3 * 2 = 17,$$

free parameters.

For question (b), $p(X_i)$ where $i = 1, 2, 3$ include three parameters. $p(X_4|X_{1:3})$ contains 2^3 free parameters. $p(X_5|X_{1:4})$ contains 2^4 free parameters. $p(X_6|X_{1:4})$ contains 2^5 free parameters. Thus we have:

$$3 + 2^3 + 2^4 + 2^5 = 59,$$

free parameters.

For question (c), the second model is easier since we only have to estimate six terms independently by counting.

10.5 Bayes nets for a rainy day

We first write down the entire probability:

$$p(V, G, R, S) = p(V) \cdot p(G) \cdot p(R|V, G) \cdot p(S|G).$$

For question (a), we have:

$$\begin{aligned} p(S = 0|V = 1) &= \frac{p(S = 0, V = 1)}{p(V = 1)} \\ &= \frac{1}{p(V = 1)} \sum_{R,G \in \{0,1\}} p(V = 1, G, R, S = 0) \\ &= 0.2\alpha\gamma + 0.8\alpha\gamma + 0.1(1 - \alpha)\beta + 0.9(1 - \alpha)\beta \\ &= \alpha\gamma + (1 - \alpha)\beta. \end{aligned}$$

For question (b), the answer is exactly the same. Since given G, S is independent from V.

For question (c), we independently estimate δ, α, γ and β:

$$\begin{aligned}\delta &= 1,\\ \alpha &= \frac{1}{3},\\ \gamma &= 1,\\ \beta &= 0.\end{aligned}$$

from counting.

10.6 Fishing nets

For question (a), we are to compute the posterior:

$$p(X_2|X_4 = \text{thin}, X_1 \sim \pi),$$

where:

$$\pi = (0.5, 0, 0, 0.5)^{\mathrm{T}}.$$

This shall be done straightforwardly:

$$\begin{aligned}p(X_2 = \text{salmon}|X_4 = \text{thin}, X_1 \sim \pi) &= \sum_{x_1,x_3} p(X_2 = \text{salmon}, X_1 = x_1, X_3 = x_3|X_4 = \text{thin})\\ &= \frac{p(X_1 = \text{winter}, X_2 = \text{salmon}, X_3 = \text{light}, X_4 = \text{thin})}{p(X_4 = \text{thin})}\\ &+ \frac{p(X_1 = \text{winter}, X_2 = \text{salmon}, X_3 = \text{medium}, X_4 = \text{thin})}{p(X_4 = \text{thin})}\\ &+ \frac{p(X_1 = \text{winter}, X_2 = \text{salmon}, X_3 = \text{dark}, X_4 = \text{thin})}{p(X_4 = \text{thin})}\\ &+ \frac{p(X_1 = \text{autumn}, X_2 = \text{salmon}, X_3 = \text{light}, X_4 = \text{thin})}{p(X_4 = \text{thin})}\\ &+ \frac{p(X_1 = \text{autumn}, X_2 = \text{salmon}, X_3 = \text{medium}, X_4 = \text{thin})}{p(X_4 = \text{thin})}\\ &+ \frac{p(X_1 = \text{autumn}, X_2 = \text{salmon}, X_3 = \text{dark}, X_4 = \text{thin})}{p(X_4 = \text{thin})}\\ &= \frac{0.5 * 0.6 * 1.7}{p(X_4 = \text{thin})}.\end{aligned}$$

While that for sea bass is:

$$p(X_2 = \text{salmon}|X_4 = \text{thin}, X_1 \sim \pi) = \frac{0.5 * 0.05 * 0.3}{p(X_4 = \text{thin})},$$

hence the fish is likely to be salmon.

For question (b),

$$\begin{aligned} p(X_1 = \text{winter}|X_3 = \text{medium}, X_4 = \text{thin}) &= \sum_{x_2} p(X_1 = \text{winter}, X_2 = x_2|X_3 = \text{medium}, X_4 = \text{thin}) \\ &= \frac{0.25 * 0.9 * 0.33 * 0.6 + 0.25 * 0.1 * 0.1 * 0.05}{p(X_3 = \text{medium}, X_4 = \text{thin})} \\ &= \frac{0.04468}{p(X_3 = \text{medium}, X_4 = \text{thin})}. \end{aligned}$$

Analogously, we have:

$$p(X_1 = \text{spring}|X_3 = \text{medium}, X_4 = \text{thin}) = \frac{0.01573}{p(X_3 = \text{medium}, X_4 = \text{thin})},$$

$$p(X_1 = \text{summer}|X_3 = \text{medium}, X_4 = \text{thin}) = \frac{0.02055}{p(X_3 = \text{medium}, X_4 = \text{thin})},$$

$$p(X_1 = \text{autumn}|X_3 = \text{medium}, X_4 = \text{thin}) = \frac{0.03985}{p(X_3 = \text{medium}, X_4 = \text{thin})}.$$

Therefore the season is likely to be winter. Intuitively, a fish that is thin and medium lightness is likely to be a salmon. And the season in which a salmon is caught is most likely to be winter.

10.7 Removing leaves in BN20 networks

For question (a), note that given $\mathbf{z}$, all components from $\mathbf{x}$ are mutually independent, hence:

$$p(\mathbf{Z}|X_1, X_2, X_4) = \frac{1}{p(X_1, X_2, X_4)} \cdot p(\mathbf{Z}) \cdot p(X_1, X_2, X_4|\mathbf{Z}),$$

according to Figure 10.16.(b), while that computed w.r.t. Figure 10.16.(a) is:

$$\begin{aligned} p(\mathbf{Z}|X_1, X_2, X_4) &= \sum_{x_3, x_5} p(\mathbf{Z}, X_3 = x_3, X_5 = x_5|X_1, X_2, X_4) \\ &= \frac{1}{p(X_1, X_2, X_4)} \cdot \sum_{x_3, x_5} p(\mathbf{Z}) \cdot p(X_1, X_2, X_4|\mathbf{Z}) p(X_3 = x_3, X_5 = x_5|\mathbf{Z}) \\ &= \frac{1}{p(X_1, X_2, X_4)} \cdot p(\mathbf{Z}) \cdot p(X_1, X_2, X_4|\mathbf{Z}) \sum_{x_3, x_5} p(X_3 = x_3, X_5 = x_5|\mathbf{Z}) \\ &= \frac{1}{p(X_1, X_2, X_4)} \cdot p(\mathbf{Z}) \cdot p(X_1, X_2, X_4|\mathbf{Z}). \end{aligned}$$

This completes the proof.

For question (b), consider the following decomposition of the joint probability:

$$\begin{aligned} p(\mathbf{X}_{\text{on}}, \mathbf{X}_{\text{off}}, \mathbf{Z}) &= p(\mathbf{Z}) \cdot \prod_{X \in \mathbf{X}_{\text{on}}} p(X|\mathbf{Z}) \cdot \prod_{Y \in \mathbf{X}_{\text{off}}} p(Y|\mathbf{Z}) \\ &= \prod_{Z \in \mathbf{Z}} p(Z) \cdot \prod_{X \in \mathbf{X}_{\text{on}}} p(X|\mathbf{Z}) \cdot \prod_{Y \in \mathbf{X}_{\text{off}}} \prod_{Z \in \mathbf{Z}} \theta_{Z,Y}, \end{aligned}$$

where $\theta_{Z,Y}$ is defined from the correlation between disease Z and symptom Y. If Z and Y are independent then $\theta_{Z,Y} = 1$.

Now:

$$p(\mathbf{X}_{\text{on}}, \mathbf{X}_{\text{off}}, \mathbf{Z}) = \prod_{Z \in \mathbf{Z}} p(Z) \left[\prod_{Y \in \mathbf{X}_{\text{off}}} \theta_{Z,Y} \right] \cdot \prod_{X \in \mathbf{X}_{\text{on}}} p(X|\mathbf{Z}).$$

By changing the prior over Z from $p(Z)$ into $p(Z) \left[\prod_{Y \in \mathbf{X}_{\text{off}}} \theta_{Z,Y} \right]$ we complete the proof.

10.8 Handling negative findings in the QMR network

Counting and conducting MLE on $p(\mathbf{d})$ is of complexity $\mathcal{O}(|\mathrm{d}|)$. For each disease, it costs $\mathcal{O}(|\mathbf{f}^-|)$ to MLE all corresponding symptoms. Hitherto we have completed the proof.

10.9 Moralization does not introduce new independence statements

To prove that:

$$\mathrm{CI}(M) \subset \mathrm{CI}(G),$$

it is sufficient to prove that:

$$\neg\mathrm{CI}(G) \subset \neg\mathrm{CI}(M),$$

where $\neg\mathrm{CI}(G)$ is the collection of denials of conditional independence from G.

A statement in $\neg\mathrm{CI}(G)$ taking form:

$$\neg\mathcal{A} \perp \mathcal{B}|\mathcal{C},$$

is tantamount to that: *there exists one unblocked path, p, from* $\mathcal{A}$ *to* $\mathcal{B}$. Consider this p in M, where the directions on the edges have vanished. The only possibility that this path in M is blocked is that p encounters a collider structure in G and the intermedium node m, or some of its descents is in $\mathcal{E}$. In this case there exists an extra edge between the parents of m in M, hence we replace the two edges connecting m and m's parents with the extra edge. This procedure ends up with another unblocked edge, hence we have the statement $\neg\text{CI}(G) \subset \neg\text{CI}(M)$ completed. This finishes the proof of the original proposition.

11 Mixture models and the EM algorithm

EM is perhaps the most important and powerful that probabilistic machine learning theory has provided. Its intuitive conciseness and analytic elegance make it the most popular method for general graph models. Apart from the properties coved in the textbook, the speed of convergence for EM is also a variable of significance. The speed of convergence for EM can be easily derived from the Hessian of the likelihood, combined with elementary matrix algebra, which has been discussed in Dempster's pioneering papers. Without an involved understanding of EM, one can hardly proceed to absorb variational inference or further stochastic learning algorithms.

Despite formal complexity, the intuition behind EM is simple and straightforward. It originates from the conjugate iterative optimization problem, which has been discussed by the optimization community for centuries. Instead of remembering all the symbolic deductions, it is this intuition that should be engraved into one's mind.

11.1 Student T as infinite mixture of Gaussian

Recall that the 1d Student-t distribution takes the form:

$$\mathrm{St}(x|\mu,\sigma^2,v)=\frac{\Gamma(\frac{v}{2}+\frac{1}{2})}{\Gamma(\frac{v}{2})}\left(\frac{1}{\pi v\sigma^2}\right)^{\frac{1}{2}}\left(1+\frac{(x-\mu)^2}{v\sigma^2}\right)^{-\frac{v+1}{2}}.$$

While the r.h.s. of (11.61) is:

$$\begin{aligned}&\int\mathcal{N}(x|\mu,\frac{\sigma^2}{z})\cdot\mathrm{Ga}\left(z|\frac{v}{2},\frac{v}{2}\right)\mathrm{d}z\\=&\int\frac{\sqrt{z}}{\sqrt{2\pi\sigma^2}}\cdot\exp\left\{-\frac{z}{2\sigma^2}(x-\mu)^2\right\}\cdot\frac{(\frac{v}{2})^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}\cdot z^{\frac{v}{2}-1}\cdot\exp\left\{-\frac{vz}{2}\right\}\mathrm{d}z\\=&\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\frac{(\frac{v}{2})^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}\int z^{\frac{v-1}{2}}\cdot\exp\left\{-\left(\frac{v}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right)z\right\}\mathrm{d}z.\end{aligned}$$

Instead of integrating analytically, we observe that the term being integrated takes the form of a Gamma density:

$$\mathrm{Ga}\left(z|\frac{v+1}{2},\frac{v}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right)=\frac{\left(\frac{v}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right)^{\frac{v+1}{2}}}{\Gamma\left(\frac{v+1}{2}\right)}\cdot z^{\frac{v-1}{2}}\cdot\exp\left\{-\left(\frac{v}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right)z\right\}.$$

Therefore the r.h.s. of (11.61) becomes:

$$\begin{aligned}\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\frac{(\frac{v}{2})^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}\cdot\frac{\Gamma\left(\frac{v+1}{2}\right)}{\left(\frac{v}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right)^{\frac{v+1}{2}}} &= \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\frac{(\frac{v}{2})^{\frac{v}{2}}}{\left(\frac{v}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right)^{\frac{v+1}{2}}}\\ &= \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\cdot(\frac{v}{2})^{-\frac{1}{2}}\left(1+\frac{(x-\mu)^2}{2v\sigma^2}\right)^{-\frac{v+1}{2}}.\end{aligned}$$

This completes the proof.

11.2 EM for mixture of Gaussians

We are to optimize the following target w.r.t. θ:

$$\begin{aligned}Q(\theta,\theta^{\text{old}}) =&\mathbb{E}_{p(\mathbf{z}|\mathcal{D},\theta^{\text{old}})}\left[\sum_{n=1}^{N}\log p(\mathbf{x}_n,\mathbf{z}_n|\theta)\right]\\ =&\sum_{n=1}^{N}\mathbb{E}_{p(\mathbf{z}|\mathcal{D},\theta^{\text{old}})}\left[\log\prod_{k=1}^{K}(\pi_k\cdot p(\mathbf{x}_n|z_k,\theta))^{z_{nk}}\right]\\ =&\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{p(\mathbf{z}|\mathcal{D},\theta^{\text{old}})}\left[z_{nk}\cdot\log(\pi_k\cdot p(\mathbf{x}_n|z_k,\theta))\right]\\ =&\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{p(\mathbf{z}|\mathcal{D},\theta^{\text{old}})}\left[z_{nk}\right]\cdot\log(\pi_k\cdot p(\mathbf{x}_n|z_k,\theta))\\ =&\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\cdot\log(\pi_k\cdot p(\mathbf{x}_n|z_k,\theta)),\end{aligned}$$

where:

$$r_{nk}=p(z_{nk}=1|\mathbf{x}_n,\theta^{\text{old}}).$$

(Recall the graphical structure of GMM model. $\mathbf{z}_n$ is the one-hot variable that encodes the belonging of sample $\mathbf{x}_n$ to the centroids.) When the base distribution $p(\mathbf{x}|\mathbf{z},\theta)$ is Gaussian, consider the terms involving μ_k and Σ_k in $Q(\theta,\theta^{\text{old}})$ first (adopting non-information prior):

$$\begin{aligned}\sum_{n=1}^{N}r_{nk}\cdot\log p(\mathbf{x}_n|z_k=1,\theta) &= \sum_{n=1}^{N}r_{nk}\cdot\log\mathcal{N}(\mathbf{x}_n|\mu_k,\sigma_k^2)\\ &= \sum_{n=1}^{N}r_{nk}\cdot\left(C-\frac{1}{2}\log|\Sigma_k|-\frac{1}{2}(\mathbf{x}_n-\mu_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{x}_n-\mu_k)\right)\\ &= \mathcal{L}(\mu_k,\Sigma_k).\end{aligned}$$

Optimizing this target w.r.t. μ_k and Σ_k is tantamount to optimizing the mean and covariance of a weighted Gaussian model, hence:

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = \sum_{n=1}^{n} r_{nk} \cdot \Sigma^{-1}(\mathbf{x}_n - \mu_k).$$

Setting it to zero yields:

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} \cdot \mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}}.$$

Finally:

$$\frac{\partial \mathcal{L}}{\partial \Lambda_k} = \sum_{n=1}^{N} r_{nk} \left(\frac{1}{2}\Lambda_k^{-1} - \frac{1}{2}(\mathbf{x} - \mu_k)(\mathbf{x} - \mu_k)^{\mathrm{T}} \right),$$

where $\Lambda_k = \Sigma_k^{-1}$. Setting it to zero yields:

$$\Sigma_k = \Lambda_k^{-1} = \frac{\sum_{n=1}^{N} r_{nk}(\mathbf{x} - \mu_k)(\mathbf{x} - \mu_k)^{\mathrm{T}}}{\sum_{n=1}^{N} r_{nk}}.$$

So far we have proven (11.114) and (11.115).

11.3 EM for mixtures of Bernoullis

For the mixture of Bernoullis model, consider K bases, from which each is a Bernoulli distribution:

$$\mathrm{Ber}(x|\theta_k) = \theta_k^{\mathbb{I}(x=1)} \cdot (1 - \theta_k)^{\mathbb{I}(x=0)}.$$

The auxiliary function, whom we are to optimize w.r.t. θ is:

$$\begin{aligned} Q(\theta, \theta^{\mathrm{old}}) &= \mathbb{E}_{p(\mathbf{z}|\mathcal{D},\theta^{\mathrm{old}})} \left[\sum_{n=1}^{N} \log p(x_n, \mathbf{z}_n|\theta) \right] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}[z_{nk}] \cdot (\log \pi_k + \mathbb{I}(x_n = 1) \log \theta_k + \mathbb{I}(x_n = 0) \log(1 - \theta_k)) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \cdot (\log \pi_k + \mathbb{I}(x_n = 1)\theta_k + \mathbb{I}(x_n = 0)(1 - \theta_k)). \end{aligned}$$

Taking differential w.r.t. θ_k

$$\frac{\partial Q}{\partial \theta_k} = \sum_{n=1}^{N} r_{nk} \cdot \left(\mathbb{I}(x_n = 1)\frac{1}{\theta_k} - \mathbb{I}(x_n = 0)\frac{1}{1 - \theta_k} \right),$$

set it to zero:

$$\theta_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbb{I}(x_n = 1)}{\sum_{n=1}^{N} r_{nk}}.$$

This is exactly (11.116) modules α-reduction.

If a Beta(α_k, β_k) prior is introduced for each base then we introduce $\alpha_k - 1$ positive samples and $\beta_k - 1$ negative samples into the computation, this is tantamount to setting $r_{nk} = 1$ for $n = N+1, \cdots, N+\alpha_k+\beta_k-2$, so:

$$\theta_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbb{I}(x_n = 1) + \alpha_k - 1}{\sum_{n=1}^{N} r_{nk} + \alpha_k + \beta_k - 2}.$$

At this point one might wonder the necessity of introducing a mixture of Bernoullis. Unlike the mixture of Gaussians, that of Bernoullis seems less convincing. Let θ denotes the weighted average of base models:

$$\theta = \sum_k \pi_k \theta_k,$$

then the variance of the mixture model remains $\theta - \theta^2$. There is no need of using a mixture of Bernoullis (regardin prediction) unless we have to explicitly model a scenario in which there has to be a mixture structure. For example, if we were told that a binary string is generated from a set of unbalanced coins where each coin has different dynamics and we are asked to tell which coin generates some specific toss. But even this scenario might lead to abnormality, considering a coin that always yields head and another that always yields tail.

11.4 EM for mixture of Student distributions

The log-likelihood for the complete data set for the mixture of Student distribution is:

$$l_c(\mathbf{x}, z_k = 1) = \pi_k \cdot \frac{\Gamma(\frac{D}{2} + \frac{v_k}{2})}{\Gamma(\frac{v_k}{2})} \cdot \frac{|\Sigma_k|^{-\frac{1}{2}}}{v_k^{\frac{D}{2}} \pi^{\frac{D}{2}}} \times \left(1 + \frac{1}{v_k}(\mathbf{x}-\mu)\Sigma_k^{-1}(\mathbf{x}-\mu)\right)^{-\frac{v_k+D}{2}}.$$

Following the similar path as exercise 11.2 and exercise 11.3, we have:

$$Q(\theta, \theta^{\text{old}}) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \cdot \left(\log \pi_k + \log\left(\frac{\Gamma(\frac{D}{2} + \frac{v_k}{2})}{\Gamma(\frac{v_k}{2})} \cdot \frac{|\Sigma_k|^{-\frac{1}{2}}}{v_k^{\frac{D}{2}} \pi^{\frac{D}{2}}} \times \left(1 + \frac{1}{v_k}(\mathbf{x}-\mu)\Sigma_k^{-1}(\mathbf{x}-\mu)\right)^{-\frac{v_k+D}{2}}\right)\right).$$

The rest is ordinary EM, whose detail symbolic forms is too complex to write down explicitly.

11.5 Gradient descent for fitting GMM

From the given (11.118) and (11.119):

$$p(\mathbf{x}|\theta) = \sum_k \pi_k \cdot \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$
$$l(\theta) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta).$$

While $r_{nk} = p(z_{nk} = 1|\mathbf{x}_n, \theta)$ is defined by (11.120).

For question (a), recall that:

$$\begin{aligned} l(\theta) &= \sum_{n=1}^{N} \log \left[\sum_{\mathbf{z}_n} p(\mathbf{x}_n, \mathbf{z}_n|\theta) \right] \\ &= \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} p(\mathbf{x}_n, z_{nk} = 1|\theta) \right] \\ &= \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \right]. \end{aligned}$$

We are now ready to taking partial gradient of l w.r.t. μ_k, which yields:

$$\begin{aligned} \frac{\partial l}{\partial \mu_k} &= \sum_{n=1}^{N} \frac{\partial}{\partial \mu_k} \log \left[\sum_{k'=1}^{K} \pi_{k'} \cdot \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'}) \right] \\ &= \sum_{n=1}^{N} \frac{\pi_k}{\sum_{k'=1}^{K} \pi_{k'} \cdot \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'})} \cdot \frac{\partial \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\partial \mu_k} \\ &= \sum_{n=1}^{N} \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \cdot \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'})} \cdot \Sigma_k^{-1}(\mathbf{x}_n - \mu_k), \end{aligned}$$

using (4.10) for the last step. Now we have arrived in (11.121).

For question (b):

$$\begin{aligned} \frac{\partial l}{\partial \pi_k} &= \sum_{n=1}^{N} \frac{\partial}{\partial \pi_k} \log \left[\sum_{k'=1}^{K} \pi_{k'} \cdot \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'}) \right] \\ &= \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \cdot \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'})}, \end{aligned}$$

For question (c), with:

$$\pi_k = \frac{\exp(w_k)}{\sum_{k'=1}^{K} \exp(w_{k'})},$$

we have:

$$\begin{aligned}
\frac{\partial l}{\partial w_k} &= \sum_j \frac{\partial l}{\partial \pi_j} \frac{\partial \pi_j}{\partial w_k} \\
&= \sum_j \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)}{\sum_{k'=1}^{K} \pi_{k'} \cdot \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'})} \cdot \pi_j (1-\pi_j)^{\mathbb{I}(j=k)} (-\pi_k)^{\mathbb{I}(j\neq k)} \\
&= \sum_{n=1}^{N} \sum_j r_{nj} (1-\pi_j)^{\mathbb{I}(j=k)} (-\pi_k)^{\mathbb{I}(j\neq k)} \\
&= \sum_{n=1}^{N} r_{nk} \cdot (1-\pi_k) + (1-r_{nk}) \cdot (-\pi_k) \\
&= \sum_{n=1}^{N} r_{nk} - \pi_k,
\end{aligned}$$

where no constant factor is missed.

For question (d), we have:

$$\begin{aligned}
\frac{\partial l}{\partial \Sigma_k} &= \sum_{n=1}^{N} \frac{\partial}{\partial \Sigma_k} \log \left[\sum_{k'=1}^{K} \pi_{k'} \cdot \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'}) \right] \\
&= \sum_{n=1}^{N} r_{nk} \left(-\frac{1}{2} \right) \left((textbf x_n \mu_k)(textbf x_n \mu_k)^{\mathrm{T}} - \Sigma_k \right),
\end{aligned}$$

during which process we need to use (4.10) and the fact:

$$\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}| \mathbf{A}^{-1},$$

for a symmetric matrix $\mathbf{A}$. Thus the optimal Σ_k takes the same form as what has been derived in exercise 11.2.

For question (e), this process is redundant since the MLE for Σ is already a positive definite matrix.

11.6 EM for a finite scale mixture of Gaussians

For question (a), we advance straightforwardly:

$$\begin{aligned} p(J_n = j, K_n = k|x_n, \theta) &= \frac{p(x_n, J_n = j, K_n = k|\theta)}{p(x_n, \theta)} \\ &= \frac{1}{A} p_j \cdot q_k \cdot \mathcal{N}(x_n|\mu_j, \sigma_k^2) \\ &= \frac{p_j \cdot q_k \cdot \mathcal{N}(x_n|\mu_j, \sigma_k^2)}{\sum_{j',k'} p_{j'} \cdot q_{k'} \cdot \mathcal{N}(x_n|\mu_{j'}, \sigma_{k'}^2)}. \end{aligned}$$

For question (b):

$$\begin{aligned} Q(\theta^{\text{new}}, \theta^{\text{old}}) &= \mathbb{E}_{p(J_n, K_n|x_n, \theta^{\text{old}})} \left[\sum_{n=1}^{N} \log p(x_n, J_n, K_k|x_n, \theta^{\text{new}}) \right] \\ &= \sum_{n=1}^{N} \mathbb{E}_{p(J_n, K_n|x_n, \theta^{\text{old}})} \left[\sum_{j=1}^{m} \sum_{k=1}^{l} z_{jk} \log \left(p_j^{\text{new}} q_k^{\text{new}} \cdot \mathcal{N}(x_n|\mu_j^{\text{new}}, \sigma_k^{2,\text{new}}) \right) \right] \\ &= \sum_{n} \sum_{j,k} \mathbb{E}_{p(J_n, K_n|x_n, \theta^{\text{old}})} \left[z_{jk} \right] \left(\log p_j^{\text{new}} + \log q_k^{\text{new}} + \log \mathcal{N}(x_n|\mu_j^{\text{new}}, \sigma_k^{2,\text{new}}) \right). \end{aligned}$$

What left is trivial calculus. Define:

$$r_{njk} = \mathbb{E}_{p(J_n, K_n|x_n, \theta^{\text{old}})} \left[z_{jk} \right] = p(J_n = j, K_n = k|x_n, \theta)$$

For question (c), we have:

$$\begin{aligned} \frac{\partial Q}{\partial \mu_{j'}^{\text{new}}} &= \sum_{n} \sum_{j,k} r_{njk} \cdot \frac{\partial}{\partial \mu_{j'}^{\text{new}}} \log \mathcal{N}(x_n|\mu_j^{\text{new}}, \sigma_j^{2,\text{new}}) \\ &= \sum_{n} \sum_{k} r_{nj'k} \cdot \frac{(x_n - \mu_{j'}^{\text{new}})}{\sigma_k^{2,\text{new}}}. \end{aligned}$$

Setting it to zero yields:

$$\mu_{j'}^{\text{new}} = \frac{\sum_n \sum_k \frac{r_{nj'k} \cdot x_n}{\sigma_k^{2,\text{new}}}}{\sum_n \sum_k \frac{r_{nj'k}}{\sigma_k^{2,\text{new}}}}.$$

11.7 Manual calculation of the M step for a GMM

For question (a), we are to optimize:

$$\sum_{n=1}^{3} \sum_{k=1}^{2} r_{nk} \left(\log \pi_i + \log \mathcal{N}(x_n|\mu_k, \sigma_k^2) \right).$$

For question (b), the new optimal assignment for π_1, π_2 is derived by differentiating the auxiliary function added with a regularizer to ensure $\pi_1 + \pi_2 = 1$:

$$\frac{\partial Q + \lambda(\pi_1 + \pi_2 - 1)}{\partial \pi_1} = \frac{\sum_{n=1}^{3} r_{n,1}}{\pi_1} + \lambda = \frac{1.4}{\pi_1} + \lambda,$$

$$\frac{\partial Q + \lambda(\pi_1 + \pi_2 - 1)}{\partial \pi_2} = \frac{\sum_{n=1}^{3} r_{n,2}}{\pi_2} + \lambda = \frac{1.6}{\pi_2} + \lambda.$$

Setting both gradients to zero ends up with:

$$\begin{aligned} \pi_1 &= \frac{7}{15}, \\ \pi_2 &= \frac{8}{15}, \\ \lambda &= -3. \end{aligned}$$

For question (c), we use the results from exercise 11.2:

$$\begin{aligned} \mu_1 &= \frac{\sum_{n=1}^{3} r_{n1} \cdot x_n}{\sum_{n=1}^{3} r_{n1}} = \frac{25}{7}, \\ \mu_2 &= \frac{\sum_{n=1}^{3} r_{n2} \cdot x_n}{\sum_{n=1}^{3} r_{n2}} = \frac{65}{4}. \end{aligned}$$

11.8 Moments of a mixture of Gaussians

For question (a), the expectation of a mixture Gaussian distribution is:

$$\begin{aligned} \mathbb{E}(\mathbf{x}) &= \int \mathbf{x} \sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \mathrm{d}\mathbf{x} \\ &= \sum_k \pi_k \left(\int \mathbf{x} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \mathrm{d}\mathbf{x} \right) \\ &= \sum_k \pi_k \mu_k. \end{aligned}$$

For question (b), recall that $\mathrm{cov}(\mathbf{x}) = \mathbb{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) - \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x})^{\mathrm{T}}$. We have:

$$\begin{aligned} \mathbb{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) &= \int \mathbf{x}\mathbf{x}^{\mathrm{T}} \sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \mathrm{d}\mathbf{x} \\ &= \sum_k \pi_k \int \mathbf{x}\mathbf{x}^{\mathrm{T}} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \mathrm{d}\mathbf{x}, \end{aligned}$$

in which:

$$\begin{aligned}\int \mathbf{x}\mathbf{x}^{\mathrm{T}}\mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k)\mathrm{d}\mathbf{x} =& \mathbb{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) \\ =& \mathrm{cov}(\mathbf{x}) + \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x})^{\mathrm{T}} \\ =& \Sigma_k + \mu_k\mu_k^{\mathrm{T}}.\end{aligned}$$

Therefore:

$$cov(\mathbf{x}) = \sum_k \pi_k(\Sigma_k + \mu_k\mu_k^{\mathrm{T}}) - \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x})^{\mathrm{T}}.$$

11.9 K-means clustering by hand

```
import math
def distance(a,b):
    return math.sqrt((a[0]-b[0])**2+(a[1]-b[1])**2)
data
    =[[1,0],[3,0],[5,0],[7,0],[9,0],[11,0],[13,0],[15,0],[17,0],

      [1,3],[3,3],[5,3],[7,3],[9,3],[11,3],[13,3],[15,3],[17,3]]

c1=[9,3]
c2=[11,3]
membership=[1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1]
for i in range(10):
    c1i=0
    c1x=0
    c1y=0
    c2i=0
    c2x=0
    c2y=0
    for j in range(len(data)):
        if (distance(data[j],c1)<=distance(data[j],c2)):
            membership[i]=1
            c1i=c1i+1
            c1x=c1x+data[j][0]
            c1y=c1y+data[j][1]
        else:
            membership[i]=2
```

```
            c2i=c2i+1
            c2x=c2x+data[j][0]
            c2x=c2x+data[j][1]
    c1[0]=c1x/c1i
    c1[1]=c1y/c1i
    c2[0]=c2x/c2i
    c2[1]=c2y/c2i
print(c1)
print(c2)
```

The result is:

```
[5.0, 1.5]
[15.5, 0.0]
```

11.10 Deriving the K-means cost function

For every term for a given k, note that:

$$\begin{aligned}\sum_{i:z_i=k}\sum_{i':z_{i'}=k}(x_i-x_{i'})^2 &= \sum_i\sum_{i'}\left[(x_i-\bar{x})-(x_{i'}-\bar{x})\right]^2\\ &= \sum_i\sum_{i'}(x_i-\bar{x})^2+(x_{i'}-\bar{x})^2-2(x_i-\bar{x})(x_{i'}-\bar{x})\\ &= 2\cdot n_k\cdot\sum_i(x_i-\bar{x})^2-2\left[\sum_i(x_i-\bar{x})\right]\left[\sum_{i'}(x_{i'}-\bar{x})\right]\\ &= 2\cdot n_k\cdot\sum_i(x_i-\bar{x})^2.\end{aligned}$$

This finishes the proof. One can also adopt (11.132)-(11.135) for the desired result.

11.11 Visible mixtures of Gaussians are in exponential family

Encode latent variable as binary code:

$$z_k=\mathbb{I}(x \text{ is generated from the } k\text{-th base distribution}),$$

then

$$p(\mathbf{z}|\theta) = \prod_{k=1}^{K} \pi_k^{z_k},$$
$$p(x|\mathbf{z},\theta) = \prod_{k=1}^{K} \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot \exp\left\{ -\frac{1}{2\sigma_k^2}(x-\mu_k)^2 \right\} \right)^{z_k},$$

where $\theta = (\pi, \mu, \sigma^2)$.

The logarithm for the joint distribution is:

$$\begin{aligned}\log p(x,\mathbf{z}|\theta) &= \log \prod_{k=1}^{K} \left(\frac{\pi_k}{\sqrt{2\pi\sigma_k^2}} \cdot \exp\left\{ -\frac{1}{2\sigma_k^2}(x-\mu_k)^2 \right\} \right)^{z_k} \\ &= \sum_{k=1}^{K} z_k \cdot \left(\log \pi_c - \frac{1}{2}\log 2\pi\sigma_k^2 - \frac{1}{2\sigma_k^2}(x-\mu_k)^2 \right).\end{aligned}$$

To reduce $p(x,\mathbf{z}|\theta)$ into the exponential family, note that $\log p(x,\mathbf{z}|\theta)$ is linearly dependent on $\mathbf{z}$ and x, hence we can rewrite:

$$\phi(x,\mathbf{z}) = (\mathbf{z}^{\mathrm{T}}, x\mathbf{z}^{\mathrm{T}}, x^2\mathbf{z}^{\mathrm{T}})^{\mathrm{T}},$$

the parameters for this form are:

$$(\log \pi_{\dots} - \frac{1}{2}\log 2\pi\sigma^2, \mu_{\dots} \odot \sigma_{\dots}^{-2}, -\frac{1}{2\sigma_{\dots}^2})^{\mathrm{T}}$$

in vector form.

For the mixture of MVN, since:

$$\log p(\mathbf{x},\mathbf{z}|\theta) = \sum_{k=1}^{K} z_k \left(\log \pi_c - \frac{D}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\mu_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{x}-\mu_k) \right)^2,$$

so the distribution is still an exponential family member, with sufficient statistics:

$$\phi(\mathbf{x},\mathbf{z}) = (\mathbf{z}, \mathbf{x}\odot\mathbf{z}, \mathbf{x}\odot\mathbf{x}\odot\mathbf{z}),$$

rearranged as a vector. ($\odot$ denotes the tensor/outer product.)

11.12 EM for robust linear regression with a Student t likelihood

Assuming that v and σ^2 are fixed, then the likelihood takes the form:

$$p(y|\mathbf{x},\mathbf{w},\sigma^2,v) = f(\sigma^2,v) \times \left(1 + \frac{(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})^2}{2\sigma^2} \right)^{-\frac{v+1}{2}}.$$

We only derive the M-step for this exercise with fixed σ^2 and v. The negative logarithm likelihood is:

$$\mathcal{L}(\mathbf{w}) \propto \sum_{i=1}^{N} \cdot \log\left(1 + \frac{(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^2}{2\sigma^2}\right).$$

Taking gradient:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} \propto \sum_{i=1}^{N} \frac{(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)\mathbf{x}_i}{2\sigma^2 + (y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^2}.$$

The dependency of the denominator on $\mathbf{w}$ makes the optimum analytically intractable. One possible solution is to randomly initialize $\mathbf{w}$ with $\mathbf{w}_0$ and conduct iteration:

$$\mathbf{w}_{t+1} = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}}{2\sigma^2 + (y_i - \mathbf{w}_t^{\mathrm{T}}\mathbf{x}_i)^2}\right)^{-1} \left(\sum_{i=1}^{n} \frac{y_i \mathbf{x}_i}{2\sigma^2 + (y_i - \mathbf{w}_t^{\mathrm{T}}\mathbf{x}_i)^2}\right).$$

11.13 EM for EB estimation of Gaussian shrinkage model

This is an example of non-mixture latent graphical model. In this case the latent variable is no longer the one-hot type, making it different from the EM forms that we have developed.

Recall that the complete likelihood for Gaussian shrinkage model is:

$$\begin{aligned} p\left(\theta, \mathcal{D}|\mu, \tau^2, \left\{\sigma_j^2\right\}_{j=1}^{D}\right) &= p(\theta|\mu, \tau^2) \cdot p\left(\mathcal{D}|\theta, \left\{\sigma_j^2\right\}_{j=1}^{D}\right) \\ &= \prod_{j=1}^{D} \left[\mathcal{N}(\theta_j|\mu, \tau^2) \prod_{i=1}^{N_j} \cdot \mathcal{N}(x_{ij}|\theta_j, \sigma_j^2)\right]. \end{aligned}$$

Taking logarithm yields:

$$\begin{aligned}
\log p\left(\theta, \mathcal{D}|\mu, \tau^2, \left\{\sigma_j^2\right\}_{j=1}^D\right) &= \sum_{j=1}^D \left[\log \mathcal{N}(\theta_j|\mu, \tau^2) + \sum_{i=1}^{N_j} \log \mathcal{N}(x_{ij}|\theta_j, \sigma_j^2)\right] \\
&= \sum_{j=1}^D \left[-\frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}(\theta_j - \mu)^2\right] \\
&\quad + \sum_{j=1}^D \sum_{i=1}^{N_j} \left[-\frac{1}{2}\log 2\pi\sigma_j^2 - \frac{1}{2\sigma_j^2}(x_{ij} - \theta_j)^2\right] \\
&= -\frac{D}{2}\log 2\pi\tau^2 - \frac{\sum_{j=1}^D (\theta_j - \mu)^2}{2\tau^2} + \sum_{j=1}^D \left[-\frac{N_j}{2}\log 2\pi\sigma_j^2\right] \\
&\quad - \sum_{j=1}^D \sum_{i=1}^{N_j} \frac{(x_{ij} - \theta_j)^2}{2\sigma_j^2}.
\end{aligned}$$

Note that $p(\theta, \mathcal{D}|\mu, \tau^2, \sigma_j^2\text{s})$ is essentially Gaussian, hence the posterior over θ can be analytically written down with (4.125) (though tedious). Hence all terms that dependent on θ in the logarithm of the complete likelihood can be estimated as moments of their posterior. This is possible since all such terms taking the form θ_j or θ_j^2. This completes the E-step.

For the M-step, this model is not different from others we have developed so far. Taking partial gradient w.r.t. μ and τ^2 and setting them to zero would yield the update rules.

11.14 EM for censored linear regression

The model for censored linear regression (non-Bayesian version) is:

$$\begin{aligned}
\epsilon_i &\leftarrow \mathcal{N}(\epsilon|0, \sigma^2), \\
z_i &= \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \epsilon_i, \\
y_i &= \min(z_i, c_i).
\end{aligned}$$

The observed variables are y_i, c_i and $\mathbf{x}_i$ and we are to estimate $\mathbf{w}$ and σ^2. The latent variable in this model is z_i. The complete likelihood is:

$$p(y_i, z_i|c_i, \mathbf{x}_i, \mathbf{w}, \sigma^2) = \mathcal{N}(y_i|\mathbf{w}^{\mathrm{T}}\mathbf{x}_i, \sigma^2),$$

if $y_i \neq c_i$, and is:

$$p(y_i, z_i | c_i, \mathbf{x}_i, \mathbf{w}, \sigma^2) = \int_{c_i}^{\infty} \mathcal{N}(z | \mathbf{w}^{\mathrm{T}} \mathbf{x}_i, \sigma^2) \mathrm{d}z,$$

if $y_i = c_i$, which implies $z_i \geq c_i$. We observe that the integral is going to appear inside the logarithm operator, so it is better to approximate this value from its moments. One possible approximation is to use (11.137) and (11.138), so when $y_i = c_i$, the first and the second moment of z_i are:

$$\mathbf{w}^{\mathrm{T}} \mathbf{x}_i + \sigma \cdot H\left(\frac{c_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}{\sigma}\right),$$
$$(\mathbf{w}^{\mathrm{T}} \mathbf{x}_i)^2 + \sigma^2 + \sigma(c_i + \mathbf{w}^{\mathrm{T}} \mathbf{x}_i) \cdot H\left(\frac{c_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}{\sigma}\right),$$

respectively. Note that this is a variant version of the E-step.

The log likelihood for the entire dataset now becomes:

$$\begin{aligned}
\log p(\mathbf{Y}, \mathbf{Z} | c_i, \mathbf{X}, \mathbf{w}, \sigma^2) &= \sum_{y_i \neq c_i} \log p(y_i, z_i | c_i, \mathbf{x}_i, \mathbf{w}, \sigma^2) + \sum_{y_i = c_i} \log p(y_i, z_i | c_i, \mathbf{x}_i, \mathbf{w}, \sigma^2) \\
&= \sum_{y_i \neq c_i} \left[-\frac{1}{2} \log 2\pi\sigma^2 - \frac{(y_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i)^2}{2\sigma^2} \right] \\
&\quad + \sum_{y_i = c_i} \left[-\frac{1}{2} \log 2\pi\sigma_i'^2 - \frac{\left(y_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i - \sigma \cdot H\left(\frac{c_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}{\sigma}\right)\right)^2}{2\sigma_i'^2} \right],
\end{aligned}$$

where $\sigma_i'^2 = \mathbb{E}[z_i^2 | z_i \geq c_i] - \mathbb{E}^2[z_i | z_i \geq c_i]$.

Finally, in the M-step, we take the partial gradient of the log likelihood w.r.t. $\mathbf{w}$ and σ^2, set them to zero in order to update them. This step is straightforward given H's gradient can be automatically solved efficiently.

11.15 Posterior mean and variance of a truncated Gaussian

Denote $A = \frac{c_i - \mu_i}{\sigma}$, for the conditional mean, by linearity:

$$\mathbb{E}[z_i | z_i \geq c_i] = \mu_i + \sigma \cdot \mathbb{E}[\epsilon_i | \epsilon_i \geq A].$$

And we have:

$$\mathbb{E}[\epsilon_i | \epsilon_i \geq A] = \frac{1}{p(\epsilon_i \geq A)} \cdot \int_A^{+\infty} \epsilon_i \cdot \mathcal{N}(\epsilon_i | 0, 1) \mathrm{d}\epsilon_i = \frac{\phi(A)}{1 - \Phi(A)} = H(A),$$

where H is defined by (11.139). (Recall the definition of the conditional expectation!) Therefore we have:

$$\mathbb{E}[z_i|z_i \geq c_i] = \mu_i + \sigma \cdot H(A).$$

Now we proceed to calculate the expectation for the squared term:

$$\mathbb{E}[z_i^2|z_i \geq c_i] = \mu_i^2 + 2\mu_i\sigma\mathbb{E}[\epsilon_i|\epsilon_i \geq A] + \sigma^2\mathbb{E}[\epsilon_i^2|\epsilon_i \geq A].$$

To evaluate $\mathbb{E}[\epsilon_i^2|\epsilon_i \geq A]$, we make use of the hint of this exercise:

$$\frac{\mathrm{d}}{\mathrm{d}w}(w \cdot \mathcal{N}(w|0,1)) = \mathcal{N}(w|0,1) - w^2 \cdot \mathcal{N}(w|0,1).$$

Hence we can solve for the following integral by parts:

$$\int_b^c w^2 \cdot \mathcal{N}(w|0,1)\mathrm{d}w = \Phi(c) - \Phi(b) - c \cdot \mathcal{N}(c|0,1) + b \cdot \mathcal{N}(b|0,1).$$

Thence:

$$\mathbb{E}[\epsilon_i^2|\epsilon_i \geq A] = \frac{1}{p(\epsilon_i \geq A)} \cdot \int_A^{+\infty} w^2 \cdot \mathcal{N}(w|0,1)\mathrm{d}w = \frac{1 - \Phi(A) + A \cdot \phi(A)}{1 - \Phi(A)}.$$

Plug it into the previous formula:

$$\begin{aligned}\mathbb{E}[z_i^2|z_i \geq c_i] &= \mu_i^2 + 2\mu_i\sigma \cdot H(A) + \sigma^2\frac{1 - \Phi(A) + A\phi(A)}{1 - \Phi(A)} \\ &= \mu_i^2 + \sigma^2 + H(A)(\sigma c_i + \sigma\mu_i).\end{aligned}$$

12 Latent linear models

PPCA is an elegent bridge between probabilistic machine learning models and classical data processing algorithms.

12.1 M-step for FA

The EM for FA, as *a useful exercise if you want to become proficient at the math*, is presented in detail as follows. As for the mixture of FAs, you can refer to *The EM Algorithm for Mixtures of Factor Analyzers*, Zoubin Gharamani, Geoffrey E.Hinton, 1996.

We begin with: (centralize $\mathbf{X}$ to cancel μ w.l.o.g):

$$\begin{aligned} p(\mathbf{z}_n) &= \mathcal{N}(\mathbf{z}|0, \mathbf{I}), \\ p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{W}, \Psi) &= \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z}, \Psi), \end{aligned}$$

where we have centralized $\mathbf{X}$ to simplify the deduction. Now we apply (4.124) and (4.125) to the two equations before, this ends up with:

$$\begin{aligned} p(\mathbf{z}_n|\mathbf{x}_n, \mathbf{W}, \Psi) &= \mathcal{N}(\mathbf{z}_n|\mathbf{m}, \Sigma), \\ \Sigma &= (\mathbf{I} + \mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W})^{-1}, \\ \mathbf{m} &= \Sigma\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n. \end{aligned}$$

The log-likelihood for the complete data set$\{\mathbf{x}, \mathbf{z}\}$ is:

$$\begin{aligned} \log \prod_{n=1}^{N} p(\mathbf{x}_n, \mathbf{z}_n|\mathbf{W}, \Psi) &= \sum_{n=1}^{N} \left[\log p(\mathbf{z}_n) + \log p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{W}, \Psi)\right] \\ &= \sum_{n=1}^{N} -\frac{\mathbf{z}_n^{\mathrm{T}}\mathbf{z}_n}{2} - \frac{N}{2}\log|\Psi| + \sum_{n=1}^{N} -\frac{1}{2}\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)^{\mathrm{T}}\Psi^{-1}\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right) \\ &= -\frac{1}{2}\left(\sum_{n=1}^{N}\mathbf{z}_n^{\mathrm{T}}\mathbf{z}_n\right) - \frac{N}{2}\log|\Psi| \\ &\quad - \frac{1}{2}\left(\sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n + \mathbf{z}_n^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W}\mathbf{z}_n - 2\mathbf{z}_n^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n\right). \end{aligned}$$

We are now ready to formulate the auxiliary function, let $\theta = (\mathbf{W}, \Psi)$:

$$\begin{aligned} Q(\theta, \theta^{\text{old}}) &= \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\theta^{\text{old}})}\left[\sum_{n=1}^{N} \log p(\mathbf{x}_n, \mathbf{z}_n|\theta)\right] \\ &= -\frac{1}{2}\sum_{n=1}^{N} \mathbb{E}[\mathbf{z}_n^{\mathrm{T}}\mathbf{z}_n] - \frac{N}{2}\log|\Psi| \\ &\quad - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{x}_n^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n + \mathbb{E}[\mathbf{z}_n^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W}\mathbf{z}_n] - 2\mathbb{E}[\mathbf{z}_n^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n\right), \end{aligned}$$

where the conditional first and second moments are:

$$\begin{aligned} \mathbb{E}[\mathbf{z}_n] &= (\mathbf{I} + \mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n = \mathbf{m}_n, \\ \mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}] &= (\mathbf{I} + \mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W})^{-1} + \mathbf{m}_n\mathbf{m}_n^{\mathrm{T}} = \Sigma + \mathbf{m}_n\mathbf{m}_n^{\mathrm{T}}. \end{aligned}$$

Therefore:

$$\begin{aligned} Q(\theta, \theta^{\text{old}}) &= -\frac{1}{2}\sum_{n=1}^{N} \mathrm{tr}\left[\Sigma + \mathbf{m}_n\mathbf{m}_n^{\mathrm{T}}\right] - \frac{N}{2}\log|\Psi| \\ &\quad - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{x}_n^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n + \mathrm{tr}\left[\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W}(\Sigma + \mathbf{m}_n\mathbf{m}_n^{\mathrm{T}})\right] - 2\mathbf{m}_n^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n\right). \end{aligned}$$

Finally, let us take partial gradient of the auxiliary function w.r.t. $\mathbf{W}$ and Ψ. We start with $\mathbf{W}$, note that Σ and $\mathbf{m}_n$ only dependents on $\mathbf{W}^{\text{old}}$, hence the first term is a constant for θ:

$$\begin{aligned} \frac{\partial Q}{\partial \mathbf{W}} &= -\frac{1}{2}\sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{W}}\mathrm{tr}\left[\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\right] - 2\cdot\frac{\partial}{\partial \mathbf{W}}\mathrm{tr}\left[\mathbf{x}_n\mathbf{m}_n^{\mathrm{T}}\Psi^{-1}\mathbf{W}\right] \\ &= -\frac{1}{2}\sum_{n=1}^{N} \mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\cdot 2\Psi^{-1}W - 2\cdot\mathbf{x}_n\mathbf{m}_n^{\mathrm{T}}\Psi^{-1}. \end{aligned}$$

Setting it to zero yields:

$$\mathbf{W} = \left(\sum_{n=1}^{N}\mathbf{x}_n\mathbf{m}_n^{\mathrm{T}}\right)\left(\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\right)^{-1}.$$

Meanwhile, let $\Lambda = \Psi^{-1}$:

$$\frac{\partial Q}{\partial \Lambda} = \frac{N}{2}\Lambda^{-1} - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} + \mathbf{W}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}} - 2\mathbf{x}_n\mathbf{m}_n^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\right).$$

Hence:

$$\Psi = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} + \mathbf{W}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}} - 2\mathbf{x}_n\mathbf{m}_n^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\right).$$

This completes the proof.

12.2 MAP estimation for the FA model

Assume that the model is exerted the prior distribution $p(\mathbf{W})$ and $p(\Psi)$. Compare with the question before, the M-step needs to be modified into:

$$\frac{\partial}{\partial \mathbf{W}}(Q + \log p(\mathbf{W})) = 0,$$
$$\frac{\partial}{\partial \Psi}(Q + \log p(\Psi)) = 0,$$

whose results dependent on the concrete form of the prior distributions.

12.3 Heuristic for assessing applicability of PCA

We derive this heuristics from an information theory's perspective. Recall that the differential entropy for a MVN is (with $\sigma_i^2 = \lambda_i^{-1}$):

$$h(\{\lambda_i\}_{i=1}^d) = \frac{1}{2}\sum_{i=1}^{d} d\log_2(2\pi\mathrm{e}) + \log_2(\sigma_i^2).$$

After PCA, the covariance for this MVN model is obatined by replacing the smallest d' variances into $\sigma^2 \to 0$, hence the difference in entropy is:

$$\Delta h(d') = \frac{1}{2}\sum_{i=1}^{d'}\log_2\frac{\sigma_i^2}{\sigma^2} = \frac{d'}{2}\log_2\lambda - \frac{1}{2}\sum_{i=1}^{d'}\log_2\lambda_i.$$

For two eigen series with the same mean $\bar{\lambda}$, it is plausible to expect that the product of the largest d' values in the series with a larger variance is larger, hence the information loss is smaller, making the PCA better regarding information compression.

12.4 Deriving the second principal component

For this exercise, minimizing J makes $\mathbf{v}_1$ and $\mathbf{v}_2$ the first and second principal component for the dataset. By definition, $z_{i,j}$ (where $j = 1, 2$) is the projection of $\mathbf{x}_i$ onto $\mathbf{v}_j$.

For question (a), $J(\mathbf{v}_1, \mathbf{v}_2)$ is the reconstruction loss measured by l_2 norm, hence we would have:

$$z_{i,2} = \mathbf{v}_2^{\mathrm{T}} \mathbf{x}_i,$$

from the physics of projection. On the other hand, we could use a more mathematical way for deduction, with:

$$J(\mathbf{v}_2, \mathbf{z}_2) = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - z_{n,1}\mathbf{v}_1 - z_{n,2}\mathbf{v}_2)^{\mathrm{T}} (\mathbf{x}_n - z_{n,1}\mathbf{v}_1 - z_{n,2}\mathbf{v}_2),$$

we have:

$$\frac{\partial J}{\partial z_{m,2}} = \frac{1}{N} (2 \cdot z_{m,2} \mathbf{v}_2^T \mathbf{v}_2 - 2\mathbf{v}_2^{\mathrm{T}} (\mathbf{x}_m - z_{m1}\mathbf{v}_1)) = 0.$$

Using the fact that $\mathbf{v}_2^{\mathrm{T}} \mathbf{v}_2 = 1$ and $\mathbf{v}_2^{\mathrm{T}} \mathbf{v}_1 = 0$, we arrive at:

$$z_{m,2} = \mathbf{v}_2^{\mathrm{T}} \mathbf{x}_m.$$

For question (b), with:

$$\tilde{J}(\mathbf{v}_2) = -\mathbf{v}_2^{\mathrm{T}} \mathbf{C} \mathbf{v}_2 + \lambda_2 (\mathbf{v}_2^{\mathrm{T}} \mathbf{v}_2) + \lambda_{12} (\mathbf{v}_2^{\mathrm{T}} \mathbf{v}_1 - 0),$$

we adopt straightforward matrix algebra:

$$\frac{\partial \tilde{J}}{\partial \mathbf{v}_2} = -2\mathbf{C}\mathbf{v}_2 + 2\lambda_2 \mathbf{v}_2 + \lambda_{12} \mathbf{v}_1,$$

where we have assumed that $\mathbf{C}$ is symmetric and semi-positive definite. The next step is to decompose $\mathbf{v}_2$ along the eigenvectors of $\mathbf{C}$ as:

$$\mathbf{v}_2 = \sum_{d=1}^{D} f_d \cdot \mathbf{u}_d,$$

where d is arranged in the inverse order of the eigenvalues y_d, i.e., $\mathbf{u}_1 = \mathbf{v}_1$, $f_1 = \lambda_1$. Taking partial gradient w.r.t. λ_2 and λ_{12} implies:

$$\sum_{d=1}^{D} f_d^2 = 1,$$

$$f_1 = 0.$$

Hence, taking the gradient w.r.t. $\mathbf{v}_2$ as zero is tantamount to write:

$$\lambda_{12}\mathbf{v}_1 = \sum_{d=1}^{D} 2f_d \cdot (y_d - \lambda_2)\mathbf{u}_d.$$

Recall that the eigenvectors for $\mathbf{C}$ are orthogonal to each other, hence we have:

$$f_1 \cdot (y_1 - \lambda_2) = \frac{\lambda_{12}}{2},$$
$$\forall d \neq 1, f_d \cdot (y_d - \lambda_2) = 0.$$

These equations tell that $\lambda_{12} = 0$, and λ_2 equals one eigenvalue of $\mathbf{C}$, y_d, whose corresponding eigenvector $\mathbf{u}_d$ is the optimal value for $\mathbf{v}_2$. (We ignore the degenerate case here, but the generalization is straightforward.) For such $\mathbf{v}_2$, the value of $\tilde{J}$ is $-y_d^2$, whose minimum is reached when $\mathbf{v}_2$ is the eigenvector corresponding to the second largest eigenvalue.

12.5 Deriving the residual error for PCA

For question (a), we have:

$$\begin{aligned}
||\mathbf{x}_n - \sum_{j=1}^{K} z_{nj}\mathbf{v}_j||^2 &= \left(\mathbf{x}_n - \sum_{j=1}^{K} z_{nj}\mathbf{v}_j\right)^{\mathrm{T}} \left(\mathbf{x}_n - \sum_{j=1}^{K} z_{nj}\mathbf{v}_j\right) \\
&= \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - 2\mathbf{x}_n^{\mathrm{T}} \sum_{j=1}^{K} z_{nj}\mathbf{v}_j + \sum_{j=1}^{K}\sum_{j'=1}^{K} z_{nj}z_{nj'}\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_{j'} \\
&= \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \sum_{j=1}^{K} z_{nj}^2,
\end{aligned}$$

which is tantamount to (12.127).

For question (b), we have:

$$\begin{aligned}
J_K &= \frac{1}{N}\sum_{n-1}^{N}\left(\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \sum_{j=1}^{K} z_{nj}^2\right) \\
&= \frac{1}{N}\sum_{n=1}^{T}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{K}\mathbf{v}_j^{\mathrm{T}}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\mathbf{v}_v \\
&= \frac{1}{N}\sum_{n=1}^{T}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \sum_{j=1}^{K}\mathbf{v}_j^{\mathrm{T}}\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\right)\mathbf{v}_j \\
&= \frac{1}{N}\sum_{n=1}^{T}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \sum_{j=1}^{K}\mathbf{v}_j^{\mathrm{T}}\mathbf{C}\mathbf{v}_j \\
&= \frac{1}{N}\sum_{n=1}^{T}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \sum_{j=1}^{K}\lambda_j.
\end{aligned}$$

For question (c), we have:

$$\begin{aligned}
J_K &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \sum_{j=1}^{K}\lambda_j \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \left(\sum_{j=1}^{d}\lambda_j - \sum_{j=K+1}^{d}\lambda_j\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \sum_{j=1}^{d}\lambda_j + \sum_{j=K+1}^{d}\lambda_j \\
&= J_d + \sum_{j=K+1}^{d}\lambda_j = \sum_{j=K+1}^{d}\lambda_j.
\end{aligned}$$

12.6 Derivation of Fisher's linear discriminant

This is but straightforward algebra:

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{w}} &= \frac{\mathbf{S}_B\mathbf{w}\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}\mathbf{S}_W\mathbf{w}}{\left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}\right)^2} \\
&= \frac{1}{\left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}\right)^2}\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}\left(\mathbf{S}_B\mathbf{w} - \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}}\mathbf{S}_W\mathbf{w}\right).
\end{aligned}$$

Note that minimizing the Fisher loss is but one option in maximizing $\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}$ while minimizing $\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}$. Modern numerical toolkits offer more choices to be explored.

12.7 PCA via successive deflation

The matrix:

$$\mathbf{I} - \mathbf{v}\mathbf{v}^{\mathrm{T}}$$

is unknown as the *projection matrix* that maps an arbitrary vector into the subspace that is orthogonal to $\mathbf{v}$, since:

$$\mathbf{v}^{\mathrm{T}}(\mathbf{I} - \mathbf{v}\mathbf{v}^{\mathrm{T}})\mathbf{x} = \mathbf{v}^{\mathrm{T}}\mathbf{x} - \mathbf{v}^{\mathrm{T}}\mathbf{x}.$$

The only requirement is that $\mathbf{v}$ is a unit vector: $\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$.

For generalization to the projection matrix to the supplementary of a multi-dimensional space spanned by a set of orthogonal and unit bases $\{\mathbf{v}_1, \cdots, \mathbf{v}_M\}$ (otherwise using the Schmidt procedure first), we only have to use:

$$\mathbf{I} - \sum_{m=1}^{M} \mathbf{v}_m \mathbf{v}_m^{\mathrm{T}}.$$

For question (a), we have:

$$\begin{aligned}
\frac{1}{N}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} &= \frac{1}{N}\left(\mathbf{I} - \mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\right)\mathbf{X}\mathbf{X}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\right)^{\mathrm{T}} \\
&= \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}} - \frac{1}{N}\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}} \\
&= \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\lambda_1\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}} \\
&= \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}} - \lambda_1\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}.
\end{aligned}$$

For question (b), let $\tilde{\mathbf{C}} = \frac{1}{N}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}$ then its principal eigenvector $\mathbf{u}$ satisfies:

$$\tilde{\mathbf{C}}\mathbf{u} = \lambda\mathbf{u}.$$

To solve this equation, expand $\mathbf{u}$ onto the eigenvectors of $\mathbf{C}$:

$$\mathbf{u} = \sum_{d=1}^{D} f_d \cdot \mathbf{v}_d,$$

then:

$$\begin{aligned}
\tilde{\mathbf{C}}\mathbf{u} &= \sum_{d=1}^{D} f_d \cdot \left(\mathbf{C} - \lambda_1\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\right)\mathbf{v}_d \\
&= \sum_{d=2}^{D} f_d \cdot \lambda_d \cdot \mathbf{v}_d.
\end{aligned}$$

Then the eigenequation becomes:

$$f_1\lambda = 0,$$
$$\forall d = 2, \cdots, D, f_d\lambda_d = f_d\lambda,$$

s.t.:

$$\sum_{d=1}^{D} f_d^2 = 1.$$

To solve for this system, we have:

$$f_1 = 0,$$
$$\lambda = \lambda_{d'}, d' \geq 2$$
$$f_{d'} = 1,$$
$$f_d = 1, d \neq d'.$$

Hence the solution is $d' = 2$, this finishes the proof.

For question (c), the procedure is roughly:

1. $[\lambda, \mathbf{u}] = f(\mathbf{C})$;
2. Save λ and $\mathbf{u}$;
3. $\mathbf{C} - = \lambda\mathbf{u}\mathbf{u}^T$;
4. Repeat.

12.8 Latent semantic indexing

Practice by yourself.

12.9 Imputation in a FA model

In this context, $\mathbf{x}_h$ and $\mathbf{x}_v$ are complementary components of the data space. Recall that:

$$p(\mathbf{x}_v, \mathbf{x}_h|\theta) = \mathcal{N}(\mathbf{x}_v, \mathbf{x}_h|\mathbf{0}, \Psi + \mathbf{W}\mathbf{W}^{\mathrm{T}}).$$

Now we use (4.69):

$$p(\mathbf{x}_h|\mathbf{x}_v, \theta) = \mathcal{N}(\mu, \Sigma),$$

where:

$$\mu = \Sigma_{hv}\Sigma_{vv}^{-1}\mathbf{x}_v,$$
$$\Sigma = \Sigma_{hh} - \Sigma_{hv}\Sigma_{vv}^{-1}\Sigma_{vh},$$

in which Σ_{hv} corresponding to the submatrix in $\Psi + \mathbf{W}\mathbf{W}^{\mathrm{T}}$.

12.10 Efficiently evaluating the PPCA density

We firstly complete the proof of Theorem 12.2.2, this can be done by plugging (12.61) back into (12.60). Recall that the MLE is where $\mathbf{W}$ satisfies (using $\frac{\partial \mathbf{A}^{-1}}{\partial \mathbf{A}} = -\mathbf{A}^{-1}\mathbf{A}^{-1}$):

$$(\mathbf{I} - \mathbf{S}\mathbf{C}^{-1})\mathbf{C}^{-1}\mathbf{W} = 0.$$

With:

$$\mathbf{W} = \mathbf{Q}\begin{pmatrix} \lambda_1 - \sigma^2 & \cdots \\ \cdots & \cdots \\ \cdots & \lambda_L - \sigma^2 \\ 0 & \cdots \\ \cdots & 0 \end{pmatrix},$$

we have:

$$\mathbf{C}^{-1} = \mathbf{Q}\begin{pmatrix} \lambda_1^{-1} & 0 & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_L^{-1} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma^2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{\sigma^2} \end{pmatrix}\mathbf{Q}^{\mathrm{T}}.$$

In which $\mathbf{Q}$ is the sigular vectors/eigen vectors for $\mathbf{S}$. This would reduce $(\mathbf{I} - \mathbf{S}\mathbf{C}^{-1})\mathbf{C}^{-1}\mathbf{W}$ to zero, hence complete the proof. For MLE of the σ^2, using the trace trick.

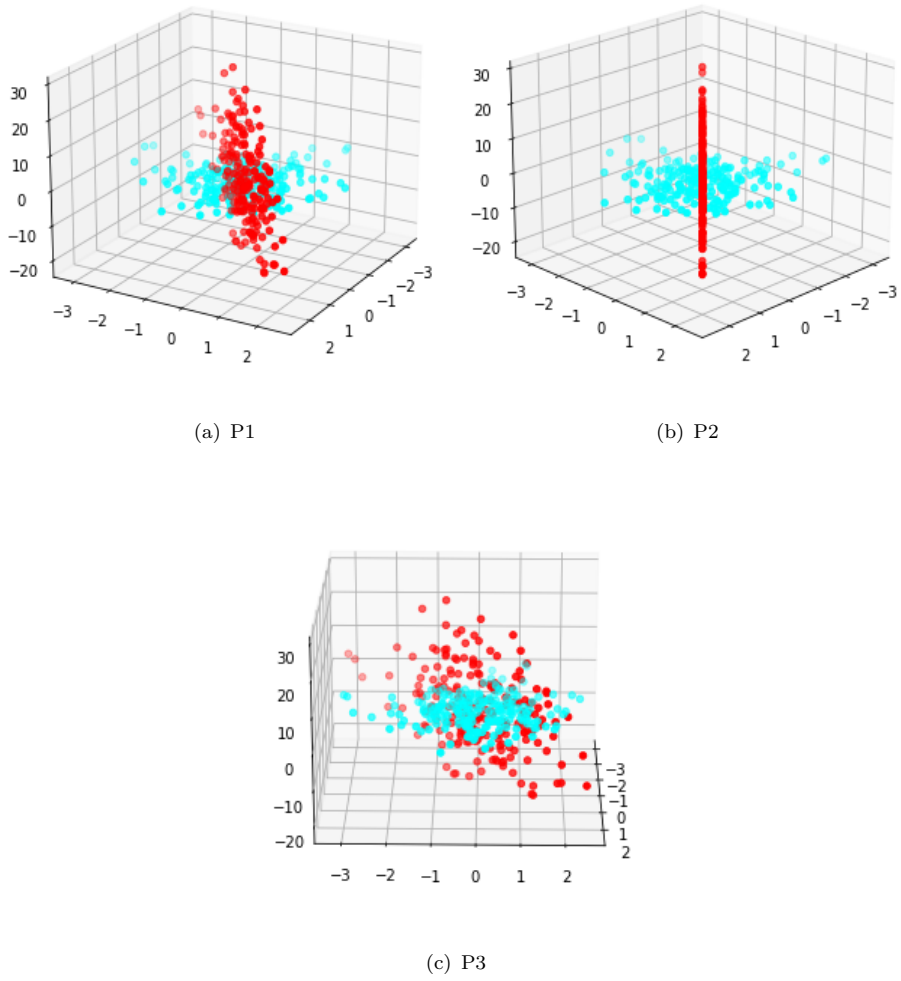
Now for $p(\mathbf{x}|\tilde{\mathbf{W}}, \tilde{\sigma}^2)$, it is a normal distribution with zero mean and covariance whose last $L - D$ components are uniformly $\tilde{\sigma}^2$.

12.11 PPCA vs FA

```
import math
import numpy as np
from numpy.linalg import eig
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d.axes3d import Axes3D
np.random.seed(520)
z1=np.random.normal(0,1,size=200)
z2=np.random.normal(0,1,size=200)
z3=np.random.normal(0,1,size=200)
x1=z1
x2=z1+0.001*z2
x3=10*z3
fig=plt.figure()
axes3d=Axes3D(fig)
axes3d.view_init(elev=20., azim=30)
axes3d.scatter(z1,z2,z3,color="cyan")
axes3d.scatter(x1,x2,x3,color="red")
Z=np.vstack((z1,z2,z3))
z1m=np.sum(z1)/200
z2m=np.sum(z2)/200
z3m=np.sum(z3)/200
zm=np.array([z1m,z2m,z3m])
for j in range(3):
    for i in range(200):
        Z[j][i]=Z[j][i]-zm[j]

X=np.vstack((x1,x2,x3))
x1m=np.sum(x1)/200
x2m=np.sum(x2)/200
x3m=np.sum(x3)/200
xm=np.array([x1m,x2m,x3m])
for j in range(3):
    for i in range(200):
        X[j][i]=X[j][i]-xm[j]
SZ=Z@Z.T/200
SX=X@X.T/200
vals,vecs=eig(SX)
print(vals)
```

The scatters are shown in the following figure: Where the red scatters are



(a) P1

(b) P2

(c) P3

Figure. 7. Exercise 12.11.

X and the cyan ones are **Z**. PCA would select dimension 3 as the principal component since it has the largest variance. PPCA would select the same if σ^2 is small enough. Otherwise, σ^2 would be estimated as approximately $\frac{100}{3}$, hence the reduced variance is larger for the first and the second dimension.

13 Sparse linear models

The techniques covered in this chapter are examples of elementary sparse models, especially linear models. In cases where we are to explore sparsity in modern machine learning models, especially deep models, we are equipped with another set of tools such as dropout, pruning and Bayesian neural networks. When the type of the hyperparameters becomes more complex, e.g., we are to decide the optimal minimal number of layers within a neural network, the sparsity remains a challenge.

13.1 Partial derivative of the RSS

For question (a), define:

$$\text{RSS}(\mathbf{w}) = \sum_{n=1}^{N} (y_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2.$$

Then we have straightforwardly:

$$\begin{aligned}
\frac{\partial}{\partial w_j}\text{RSS}(\mathbf{w}) &= \sum_{n=1}^{N} 2 \cdot (y_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)(-x_{nj}) \\
&= -\sum_{n=1}^{N} 2 \cdot (x_{nj}y_n - x_{nj}\sum_{i=1}^{D} w_i x_{ni}) \\
&= -\sum_{n=1}^{N} 2(x_{nj}y_n - x_{nj}\sum_{i\neq j}^{D} w_i x_{ni} - x_{nj}^2 w_j).
\end{aligned}$$

From which we observe that w_j's coefficient is:

$$a_j = 2\sum_{n=1}^{N} x_{nj}^2,$$

while the rest irrelevent terms can be absorbed into:

$$c_j = 2\sum_{n=1}^{N} x_{nj}(y_n - \mathbf{w}_{-j}^{\mathrm{T}}\mathbf{x}_{n,-j}).$$

The optimal value for w_j is:

$$\hat{w}_j = \frac{c_j}{a_j}.$$

For question (b), (13.184) is obvious by plugging the definition of $\mathbf{r}_k$ into (13.182)-(13.183) and the expression for $\hat{w}_j$.

13.2 Derivation of M-step for EB for linear regression

We give the EM for Automatic Relevance Determination(ARD) for the linear regression scene, the model is defined by:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X},\mathbf{w},\beta) &= \mathcal{N}(\mathbf{Y}|\mathbf{X}^{\mathrm{T}}\mathbf{w},\beta^{-1}\mathbf{I}_N), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1}), \\ \mathbf{A} &= \begin{pmatrix} \alpha_1 & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \alpha_D \end{pmatrix}. \end{aligned}$$

In which the latent variables are $\mathbf{w}$, and it is the hyperparameters $\mathbf{A}$ that requires estimation.

During the E-step, we are to estimate the expectation of $\mathbf{w}$, conditioned on $\mathbf{X}$ and $\mathbf{Y}$. Recall (4.125):

$$p(\mathbf{w}|\mathbf{Y},\mathbf{X},\beta,\mathbf{A}) = \mathcal{N}(\mathbf{w}|\mu_{\mathrm{E}},\Sigma_{\mathrm{E}}),$$

where:

$$\begin{aligned} \Sigma_{\mathrm{E}} &= \left(\mathbf{A}+\beta\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}, \\ \mu_{\mathrm{E}} &= \beta\Sigma_{\mathrm{E}}\mathbf{X}\mathbf{Y}. \end{aligned}$$

We are now ready to write down the logarithm of the complete likelihood:

$$\begin{aligned} p(\mathbf{Y},\mathbf{w}|\mathbf{X},\beta,\mathbf{A}) &= \mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A})\cdot\mathcal{N}(\mathbf{Y}|\mathbf{X}^{\mathrm{T}}\mathbf{w},\beta^{-1}\mathbf{I}_N) \\ &\propto |\mathbf{A}|^{\frac{1}{2}}\beta^{\frac{N}{2}}\exp\left\{-\frac{\beta}{2}\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right\}. \end{aligned}$$

Hence the auxiliary function is (let $\theta=(\beta,\mathbf{A})$):

$$\begin{aligned} Q(\theta,\theta^{\mathrm{old}}) &= \mathbb{E}\left[-\frac{1}{2}\log|\mathbf{A}|+\frac{N}{2}\log\beta-\frac{\beta}{2}\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right] \\ &= -\frac{1}{2}\log|\mathbf{A}|+\frac{N}{2}\log\beta-\frac{\beta}{2}\mathbb{E}\left[\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)\right]-\frac{1}{2}\mathbb{E}\left[\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right]. \end{aligned}$$

The dependence of Q on $\mathbf{A}$ is through:

$$-\frac{1}{2}\log|\mathbf{A}|-\frac{1}{2}\mathrm{tr}\left(\mathbf{A}\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\right) = -\frac{1}{2}\log|\mathbf{A}|-\frac{1}{2}\mathrm{tr}\left(\mathbf{A}(\Sigma_{\mathrm{E}}+\mu_{\mathrm{E}}\mu_{\mathrm{E}}^{\mathrm{T}})\right).$$

Hence:

$$\frac{\partial Q}{\partial\alpha_j} = \frac{1}{2\alpha_j}-\frac{1}{2}(\Sigma_{\mathrm{E}}+\mu_{\mathrm{E}}\mu_{\mathrm{E}}^{\mathrm{T}})_{jj}.$$

Using a conjugate onto α_j yields (13.166).

Finally, the dependence of Q on β is through:

$$\frac{N}{2}\log\beta - \frac{\beta}{2}\mathbb{E}\left[\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)\right],$$

where:

$$\begin{aligned}\mathbb{E}\left[\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{X}^{\mathrm{T}}\mathbf{w}\right)\right] &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - 2\mathbb{E}[\mathbf{w}]^{\mathrm{T}}\mathbf{X}\mathbf{Y} + \mathrm{tr}(\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]) \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - 2\mu_{\mathrm{E}}\mathbf{X}\mathbf{Y} + \mathrm{tr}(\mathbf{X}\mathbf{X}^{\mathrm{T}}[\Sigma_{\mathrm{E}} + \mu_{\mathrm{E}}\mu_{\mathrm{E}}^{\mathrm{T}}]).\end{aligned}$$

Hence we have:

$$\beta = \frac{N}{\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - 2\mu_{\mathrm{E}}\mathbf{X}\mathbf{Y} + \mathrm{tr}(\mathbf{X}\mathbf{X}^{\mathrm{T}}[\Sigma_{\mathrm{E}} + \mu_{\mathrm{E}}\mu_{\mathrm{E}}^{\mathrm{T}}])}.$$

Adding a Gamma prior results in (13.168).

13.3 Derivation of fixed point updates for EB for linear regression

Instead of EM, this method directly optimize the posterior probability, whose negative logarithm is:

$$l(\alpha,\beta) = -\log p(\mathbf{Y}|\mathbf{X},\alpha,\beta) + \sum_j (a\cdot\log\alpha_j - b\cdot\alpha_j) + c\cdot\log\beta - d\cdot\beta.$$

By (4.126), marginalizing out $\mathbf{w}$ yields:

$$p(\mathbf{Y}|\mathbf{X},\alpha,\beta) = \mathcal{N}(\mathbf{Y}|\mathbf{0},\Sigma_{\mathbf{Y}}),$$

where:

$$\Sigma_{\mathbf{Y}} = \beta^{-1}\mathbf{I}_N + \mathbf{X}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{X}.$$

With:

$$l(\alpha,\beta) = \frac{1}{2}\log|\Sigma_{\mathbf{Y}}| + \frac{1}{2}\mathbf{Y}^{\mathrm{T}}\Sigma_{\mathbf{Y}}^{-1}\mathbf{Y} + \sum_j (a\cdot\log\alpha_j - b\cdot\alpha_j) + c\cdot\log\beta - d\cdot\beta.$$

We have (using the matrix inverse lemma):

$$\begin{aligned}\frac{\partial\mathbf{Y}^{\mathrm{T}}\Sigma_{\mathbf{Y}}^{-1}\mathbf{Y}}{\partial\alpha_j} &= \frac{\partial}{\partial\alpha_j}\beta^{-2}(\mathbf{X}\mathbf{Y})^{\mathrm{T}}\Sigma_{\mathrm{E}}\mathbf{X}\mathbf{Y} \\ &= \frac{1}{2}\mu_{\mathrm{E},j}\mu_{\mathrm{E},j},\end{aligned}$$

on the other hand, to find:

$$\frac{\partial \log |\Sigma_{\mathrm{Y}}|}{\partial \alpha_j},$$

note that:

$$\Sigma_{\mathbf{Y},m,n} = \sum_{j=1}^{D} x_{mj} x_{nj} \alpha_j^{-1} + \beta^{-1} \mathbb{I}(m = n).$$

Thus the term $\frac{\partial \log |\Sigma_{\mathrm{Y}}|}{\partial \alpha_j}$ can be boiled down to $\gamma \cdot \alpha_j^{-1}$, hence the update is:

$$\alpha_j = \frac{2\gamma + 2a}{\mu_{\mathrm{E},j}^2 + 2b}.$$

For β, the procedure is similar.

13.4 Marginal likelihood for linear regression

This is straightforward algebra:

$$\begin{aligned} p(\mathcal{D}|\gamma) &= \int \int \mathcal{N}(\mathbf{Y}|\mathbf{X}^{\mathrm{T}}\mathbf{w}, \sigma^2\mathbf{I}) \mathcal{N}(\mathbf{w}|\mathbf{0}, \sigma^2\Sigma) p(\sigma^2) \mathrm{d}\mathbf{w} \mathrm{d}\sigma^2 \\ &\propto \int \int (\sigma^2)^{-\frac{N}{2} - \frac{D}{2} - 1} \exp\left\{ -\frac{1}{2\sigma^2} \left[\mathbf{w}^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \Sigma^{-1}) \mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{Y} + \mathbf{Y}^{\mathrm{T}}\mathbf{Y} \right] \right\} \mathrm{d}\mathbf{w} \mathrm{d}\sigma^2. \end{aligned}$$

We firstly integrate out $\mathbf{w}$:

$$\int \exp\left\{ -\frac{1}{2\sigma^2} \left[\mathbf{w}^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \Sigma^{-1}) \mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{Y} + \mathbf{Y}^{\mathrm{T}}\mathbf{Y} \right] \right\} \mathrm{d}\mathbf{w} \propto \left[\sigma^2 (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \Sigma^{-1}) \right]^{-\frac{1}{2}} \exp\left\{ -\frac{S(\gamma)}{2\sigma^2} \right\}.$$

(It seems that two factors are lost in S given by (13.17). Finally, we intergrate out σ^2:

$$\int (\sigma^2)^{-\frac{N+a}{2}} \exp\left\{ -\frac{S(\gamma)}{2\sigma^2} \right\} \mathrm{d}\sigma^2,$$

what left is tedious calculus. The plugging-in of the g-prior is trivial.

13.5 Reducing elastic net to lasso

Expanding the l.h.s. of (13.196) yields:

$$\begin{aligned} J_1(c\mathbf{w}) =& (\mathbf{y} - c\mathbf{X}\mathbf{w})^{\mathrm{T}} (\mathbf{y} - c\mathbf{X}\mathbf{w}) + c^2 \lambda_2 \mathbf{w}^{\mathrm{T}}\mathbf{w} + \lambda_1 |\mathbf{w}|_1 \\ =& \mathbf{y}^{\mathrm{T}}\mathbf{y} + c^2 \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} + c^2\lambda_2 \mathbf{w}^{\mathrm{T}}\mathbf{w} + \lambda_1 |\mathbf{w}|_1. \end{aligned}$$

While for its r.h.s.:

$$\begin{aligned} J_2(\mathbf{w}) = &\begin{pmatrix} \mathbf{y} - c\mathbf{X}\mathbf{w} \\ -c\sqrt{\lambda_2}\mathbf{w} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mathbf{y} - c\mathbf{X}\mathbf{w} \\ -c\sqrt{\lambda_2}\mathbf{w} \end{pmatrix} + c\lambda_1|\mathbf{w}|_1 \\ =&(\mathbf{y} - c\mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y} - c\mathbf{X}\mathbf{w}) + c^2\lambda_2\mathbf{w}^{\mathrm{T}}\mathbf{w} + c\lambda_1|\mathbf{w}|_1 \\ =&\mathbf{y}^{\mathrm{T}}\mathbf{y} + c^2\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} + c^2\lambda_2\mathbf{w}^{\mathrm{T}}\mathbf{w} + c\lambda_1|\mathbf{w}|_1. \end{aligned}$$

Hence (13.192) and (13.193) are equal, this proceeds to prove (13.195).

This shows elastic net regularization, which pick a regularing term as a linear combination of l_1 andl_0 equals a lasso one. (The design matrix used in this exercise collets data as rows.)

13.6 Shrinkage in linear regression

For the ordinary least square, the loss is defined as:

$$\mathrm{RSS}(\mathbf{w}) = (\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\mathbf{w})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\mathbf{w}).$$

Since $\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{I}$:

$$\mathrm{RSS}(\mathbf{w}) = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \mathbf{w}^{\mathrm{T}}\mathbf{w} - 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{w}.$$

Take its derivative w.r.t. w_k, where we note that all D weights has been decoupled in the RSS:

$$\frac{\partial}{\partial w_k}\mathrm{RSS}(\mathbf{w}) = 2w_k - 2\sum_{n=1}^{N} y_n x_{nk}.$$

Therefore we ends up with:

$$\hat{w}_k^{\mathrm{OLS}} = \sum_{n=1}^{N} y_n x_{nk} = \frac{c_k}{2}.$$

For the ridge regression:

$$\mathrm{RSS}(\mathbf{w}) = (\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\mathbf{w})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\mathbf{w}) + \lambda_2\mathbf{w}^{\mathrm{T}}\mathbf{w}.$$

Take its derivative and set it to zero:

$$(2 + 2\lambda_2)w_k = 2\sum_{n=1}^{N} y_n x_{nk}.$$

Thus

$$\hat{w}_k^{\text{ridge}} = \frac{\sum_{n=1}^N y_n x_{nk}}{1+\lambda_2} = \frac{c_k}{2(1+\lambda_2)}.$$

Finally, recall that

$$\hat{w}_k^{\text{lasso}} = \text{sign}(\hat{w}_k^{\text{OLS}})(|\hat{w}_k^{\text{OLS}}| - \frac{\lambda_1}{2})_+.$$

Observe Figure 13.24, it is easy to address the black line as OLS, the gray one Ridge and the dotted one lasso. Obviously $\lambda_1 = \lambda_2 = 1$. It is noticeable that ridge cause a shrinkage to horizontal axis while lasso cause a sharp shrinkage to zero under certain threshold.

13.7 Prior for the Bernoulli rate parameter in the spike and slab model

Recall that the prior takes the following decomposition:

$$p(\gamma|\alpha_1,\alpha_2) = \prod_{d=1}^D p(\gamma_d|\alpha_1,\alpha_2) = \prod_{d=1}^D p(\gamma_d|\pi_d)\cdot \text{Beta}(\pi_d|\alpha_1,\alpha_2).$$

We now integrate out π_d for the d-th parameter:

$$\begin{aligned}
p(\gamma_d|\alpha_1,\alpha_2) =& \frac{1}{B(\alpha_1,\alpha_2)}\int \pi_d^{\gamma_d}(1-\pi_d)^{(1-\gamma_d)}\pi_d^{\alpha_1-1}(1-\pi_d)^{\alpha_2-1}\text{d}\pi_d \\
=& \frac{1}{B(\alpha_1,\alpha_2)}\int \pi_d^{\alpha_1+\gamma_d-1}(1-\pi_d)^{\alpha_2+1-\gamma_d-1}\text{d}\pi_d \\
=& \frac{B(\alpha_1+\gamma_d,\alpha_2+1-\gamma_d)}{B(\alpha_1,\alpha_2)} = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\cdot\frac{\Gamma(\alpha_1+\gamma_d)\Gamma(\alpha_2+1-\gamma_d)}{\Gamma(\alpha_1+\alpha_2+1)}.
\end{aligned}$$

Therefore(N_1 marks the number of activa parameters in γ):

$$\begin{aligned}
p(\gamma|\alpha_1,\alpha_2) =& \binom{N}{N_1}\cdot\frac{\Gamma(\alpha_1+\alpha_2)^N}{\Gamma(\alpha_1)^N\Gamma(\alpha_2)^N}\cdot\frac{\Gamma(\alpha_1+1)^{N_1}\Gamma(\alpha_2+1)^{N-N_1}}{\Gamma(\alpha_1+\alpha_2+1)^N} \\
=& \binom{N}{N_1}\cdot\frac{\alpha_1^{N_1}\alpha_2^{N-N_1}}{(\alpha_1+\alpha_2)^N}.
\end{aligned}$$

Hence $p(\gamma)$ is this case is a binomial distribution.

13.8 Deriving E step for GSM prior

The hints contains typoes and are unnecessary. By definition, we are to calculate:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\tau_j^2}|w_j\right] &= \int \frac{1}{\tau_j^2} p(\tau_j^2|w_j)\mathrm{d}\tau_j^2 = \int \frac{1}{\tau_j^2}\frac{p(w_j|\tau_j^2)p(\tau_j^2)}{p(w_j)}\mathrm{d}\tau_j^2 \\ &= \frac{1}{p(w_j)}\int \frac{1}{\tau_j^2}\mathcal{N}(w_j|0,\tau_j^2)p(\tau_j^2)\mathrm{d}\tau_j^2 \\ &= \int \frac{1}{\sqrt{2\pi}}(\tau_j^2)^{-1.5}\exp(-\frac{w_j^2}{2\tau_j^2})p(\tau_j^2)\mathrm{d}\tau_j^2.\end{aligned}$$

While we also have:

$$p'(w_j) = \int \frac{1}{\sqrt{2\pi}}(\tau_j^2)^{-1.5}\exp(-\frac{w_j^2}{2\tau_j^2})p(\tau_j^2)\cdot w_j\mathrm{d}\tau_j^2.$$

Plugging both side into (13.197) finishes the proof.

13.9 EM for sparse probit regression with Laplace prior

The ordinary Probit regression involves no latent variable. Introducing Laplace prior for the linear weight $\mathbf{w}$ results in its lasso version. Since Laplace distribution is a continuous mixture of Gaussian according to (13.86), a latent variable τ^2 with the same dimension as $\mathbf{w}$ is introduced. For each component w_j of $\mathbf{w}$, there is a corresponding latent variable τ_j^2 to guide its variance. The PGM for this Probit regression looks like:

$$\gamma \to \tau^2 \to \mathbf{w} \to \mathbf{y} \leftarrow \mathbf{X}.$$

The joint distribution is:

$$p(\tau^2, \mathbf{w}, \mathbf{y}|\mathbf{X}, \gamma) = \left(\prod_{d=1}^{D} p(\tau_d^2|\gamma)p(w_d|\tau_d^2)\right)\cdot\left(\prod_{n=1}^{N}\Phi(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^{y_n}(1-\Phi(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n))^{1-y_n}\right),$$

where Φ is the c.d.f. for a unit Gaussian. According to (13.86):

$$\begin{aligned}p(\tau^2|\gamma) &= \mathrm{Ga}(\tau_d^2|1, \frac{\gamma^2}{2}), \\ p(w_d|\tau_d^2) &= N(w_d|0,\tau_d^2).\end{aligned}$$

Hence:

$$p(\tau^2, \mathbf{w}, \mathbf{y}|\mathbf{X}, \gamma) \propto \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left(\gamma^2\tau_d^2 + \frac{w_d^2}{\tau_d^2}\right)\right\} \cdot \prod_{d=1}^{D}\frac{1}{\tau_d} \cdot \prod_{n=1}^{N}\Phi(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^{y_n}(1-\Phi(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n))^{1-y_n}.$$

To build the auxiliary function, we assumed $\mathbf{w}$ as the parameter to be estimated and τ^2 as latent variable, thus:

$$Q(\mathbf{w}, \mathbf{w}^{\mathrm{old}}) = \mathbb{E}_{p(\tau^2|\mathbf{w}^{\mathrm{old}},\mathbf{y},\mathbf{X},\gamma)}[\log p(\tau^2, \mathbf{w}, \mathbf{y}|\mathbf{X}, \gamma)].$$

We now extract terms involving $\mathbf{w}$ from $\log p(\tau^2, \mathbf{w}, \mathbf{y}|\mathbf{X}, \gamma)$:

$$Q(\mathbf{w}, \mathbf{w}^{\mathrm{old}}) = c - \frac{1}{2}\sum_{d=1}^{D}\frac{w_d^2}{\tau_d^2} + \sum_{n=1}^{N} y_n \log\Phi(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) + (1-y_n)(1-\Phi(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)).$$

Thus we only need to calculate the conditional expectation:

$$\mathbb{E}\left[\frac{1}{\tau_d^2}|\mathbf{w}^{old}\right]$$

for the E-step. Whose result is already given in exercise 13.8. The M-step is the same as Gaussian-prior Probit regression and hence is omitted.

13.10 GSM representation of group lasso

Note that the prior over $\mathbf{w}_g$ takes the form:

$$u_g^{\frac{1}{2}}(\rho u_g)^{-\frac{1}{2}}\exp(-\rho u_g).$$

Therefore its logarithm takes the form:

$$c - \|\mathbf{w}_g\|_2,$$

where c is a term that is independent from $\mathbf{w}_g$ and the data. Combining this term with the ordinary data-dependent NLL yields the loss function of the group lasso.

13.11 Projected gradient descent for l1 regularized least squares

Generally, we take the gradient of $\mathbf{w}$ and optimize. When some constraint on $\mathbf{w}$ is broken by the gradient descent, the increment is moderated so that the constraint remains valid. For the following loss function:

$$\min_{\mathbf{w}} \left\{ \mathrm{NLL}(\mathbf{w}) + \lambda ||\mathbf{w}||_1 \right\},$$

consider under a linear regression context:

$$\mathrm{NLL}(\mathbf{w}) = \frac{1}{2} ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2.$$

For $\lambda \|\mathbf{w}\|_1$ which is not differentiate, it is suggest:

$$\mathbf{w} = \mathbf{u} - \mathbf{v},$$

where

$$u_i = (x_i)_+ = \max\{0, x_i\},$$
$$v_i = (-x_i)_+ = \max\{0, -x_i\}.$$

With $\mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}$, we have:

$$\|\mathbf{w}\|_1 = \mathbf{1}_n^{\mathrm{T}} \mathbf{u} + \mathbf{1}_n^{\mathrm{T}} \mathbf{v}.$$

Hence the original problem is translated into:

$$\min_{\mathbf{w}} \left\{ \frac{1}{2} ||\mathbf{y} - \mathbf{X}(\mathbf{u} - \mathbf{v})||_2^2 + \lambda \mathbf{1}_n^{\mathrm{T}} \mathbf{u} + \lambda \mathbf{1}_n^{\mathrm{T}} \mathbf{v} \right\}.$$
$$\text{s.t. } \mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}.$$

Denote:

$$\mathbf{z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix},$$

then we can rewrite the original target to be optimized into:

$$\min_{\mathbf{z}} \left\{ f(\mathbf{z}) = \mathbf{c}^{\mathrm{T}} \mathbf{z} + \frac{1}{2} \mathbf{z}^{\mathrm{T}} \mathbf{A} \mathbf{z} \right\},$$
$$\text{s.t. } \mathbf{z} \geq \mathbf{0},$$

where:

$$\mathbf{c} = \begin{pmatrix} \lambda \mathbf{1}_n - \mathbf{y}\mathbf{X} \\ \lambda \mathbf{1}_n + \mathbf{y}\mathbf{X} \end{pmatrix},$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & -\mathbf{X}^{\mathrm{T}}\mathbf{X} \\ -\mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{X} \end{pmatrix}.$$

Now we have changed the problem to a quadratic problem with a simple bound constraint. The gradient is given by:

$$\nabla f(\mathbf{z}) = \mathbf{c} + \mathbf{A}\mathbf{z}.$$

An ordinary gradient descent step is:

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha \nabla f(\mathbf{z}^k).$$

For projected case, take $\mathbf{g}^k$:

$$\mathbf{g}_i^k = \min \left\{ \mathbf{z}_i^k, \alpha \nabla f(\mathbf{z}^k)_i \right\}.$$

And:

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \mathbf{g}^k,$$

hence $\mathbf{z}$ is constrained as a legal weight candidate.

The original paper suggest more delicate method to moderate the learning rate, refer to *Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems, Mario A.T.Figueiredo.*

13.12 Subderivative of the hinge loss function

When $\theta < 1$, we have:

$$\partial f(\theta) = \{-1\}.$$

When $\theta = 1$, we have:

$$\partial f(\theta) = [-1, 0],$$

as for $\theta > 1$:

$$\partial f(\theta) = \{0\}.$$

13.13 Lower bounds to convex functions

Since f is a convex function, then any hyperplane that is tangent to an arbitrary point $\mathbf{x}' \in \text{dom}(f)$ has the entire f above it. Denote the normal of the hyperplane at $\mathbf{x}'$ by $\mathbf{n}'_{\mathbf{x}}$, then the hyperplane intersect with $\mathbf{x}$ at $(\mathbf{x}, \phi'_{\mathbf{x}})$ with:

$$(\mathbf{x}' - \mathbf{x}, f(\mathbf{x}') - \phi'_{\mathbf{x}}) \cdot \mathbf{n}'_{\mathbf{x}} = 0.$$

And we have:

$$f(\mathbf{x}) \geq \phi'_{\mathbf{x}}.$$

This assertation holds for arbitrary $\mathbf{x}'$, hence:

$$f(\mathbf{x}) \geq \sup_{\mathbf{x}'}(\phi'_{\mathbf{x}}).$$

For more details, refer to *Rigorous Affine Lower Bound Functions for Multivariate Polynomials and Their Use in Global Optimisation.*

14 Kernels

15 Gaussian processes

15.1 Reproducing property

We denote $\kappa(\mathbf{x}_1, \mathbf{x})$ by $f(\mathbf{x})$ and $\kappa(\mathbf{x}_2, \mathbf{x})$ by $g(\mathbf{x})$. From definition:

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} f_i \phi(\mathbf{x})$$

$$\kappa(\mathbf{x}_1, \mathbf{x}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}_1)\phi_i(\mathbf{x})$$

Since $\mathbf{x}$ can be chosen arbitrarily, we have the properties hold(the one for g is obtained similarly):

$$f_i = \lambda_i \phi_i(\mathbf{x}_1)$$

$$g_i = \lambda_i \phi_i(\mathbf{x}_2)$$

Therefore:

$$\begin{aligned} < \kappa(\mathbf{x}_1, .), \kappa(\mathbf{x}_2, .) > &= < f, g > \\ &= \sum_{i=1}^{\infty} \frac{f_i g_i}{\lambda_i} \\ &= \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}_1)\phi_i(\mathbf{x}_2) \\ &= \kappa(\mathbf{x}_1, \mathbf{x}_2) \end{aligned}$$

16 Adaptive basis function models

16.1 Nonlinear regression for inverse dynamics

Practise by yourself.

17 Markov and hidden Markov models

17.1 Derivation of Q function for HMM

Firstly, we estimate the distribution of $\mathbf{z}_{1:T}$ w.r.t θ^{old}, for auxiliay function, we are to calculate the log-likelihood w.r.t θ and $\mathbf{z}_{1:T}$.

$$\begin{aligned}
Q(\theta, \theta^{old}) =& \mathbb{E}_{p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T},\theta^{old})}[\log p(\mathbf{z}_{1:T}, \mathbf{x}_{1:T}|\theta)] \\
=& \mathbb{E}_p[\log \left\{ \prod_{i=1}^{N} \left\{ p(z_{i,1}|\pi) \prod_{t=2}^{T_i} p(z_{i,t}|z_{i,t-1}, \mathbf{A}) \prod_{t=1}^{T_i} p(x_{i,t}|z_{i,t}, \mathbf{B}) \right\} \right\}] \\
=& \mathbb{E}_p[\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}[z_{i,1} = k] \log \pi_k + \sum_{i=1}^{N} \sum_{t=2}^{T_i} \sum_{j=1}^{K} \sum_{k=1}^{K} \mathbb{I}[z_{i,t} = k, z_{i,t-1} = j] \log \mathbf{A}(j,k) \\
&+ \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{k=1}^{K} \mathbb{I}[z_{i,t} = k] \log p(x_{i,t}|z_{i,t} = k, \mathbf{B})]
\end{aligned}$$

Further we have 17.98, 17.99, 17.100, using the definition of expectation yields to 17.97.

17.2 Two filter approach to smoothing in HMMs

For $r_t(i) = p(z_t = i|x_{t+1:T})$, we have:

$$\begin{aligned}
p(z_t = i|x_{t+1:T}) &= \sum_j p(z_t = i, z_{t+1} = j|x_{t+1:T}) \\
&= \sum_j p(z_{t+1} = j|x_{t+1:T}) p(z_t = i|z_{t+1} = j, x_{t+1:T}) \\
&= \sum_j p(z_{t+1} = j|x_{t+1:T}) p(z_t = i|z_{t+1} = j) \\
&= \sum_j p(z_{t+1} = j|x_{t+1:T}) \Psi^-(j,i)
\end{aligned}$$

Where Ψ^- denotes the transform matrix in an inverse sense, we further have:

$$\begin{aligned}
p(z_{t+1} = j|x_{t+1:T}) =& p(z_{t+1} = j|x_{t+1}, x_{t+2:T}) \\
\propto& p(z_{t+1} = j, x_{t+1}, x_{t+2:T}) \\
=& p(x_{t+2:T}) p(z_{t+1} = j|x_{t+2:T}) p(x_{t+1}|z_{t+1} = j, x_{t+2:T}) \\
\propto& r_{t+1}(j) \phi_{t+1}(j)
\end{aligned}$$

Therefore we can calculate $r_t(i)$ recursively:

$$r_t(i) \propto \sum_j r_{t+1}(j)\phi_{t+1}(j)\Psi^-(j,i)$$

And initial element $p(z_T)$ is given by $\prod_T(i)$.

To rewrite $\gamma_t(i)$ in terms of new factors:

$$\begin{aligned}
\gamma_t(i) \propto & p(z_t = i|x_{1:T}) \\
\propto & p(z_t = i, x_{1:T}) \\
= & p(z_t = i)p(x_{1:T}|z_t = i) \\
= & p(z_t = i)p(x_{1:t}|z_t = i)p(x_{t+1:T}|z_t = i, x_{1:t}) \\
= & p(z_t = i)p(x_{1:t}|z_t = i)p(x_{t+1:T}|z_t = i) \\
= & \frac{1}{p(z_t = i)}p(x_{1:t}, z_t = i)p(x_{t+1:T}, z_t = i) \\
\propto & \frac{1}{p(z_t = i)}p(z_t = i|x_{1:t})p(z_t = i|x_{t+1:T}) \\
= & \frac{\alpha_t(i) \cdot r_t(i)}{\prod_t(i)}
\end{aligned}$$

17.3 EM for HMMs with mixture of Gaussian observations

Using mixture of Gaussians as the emission distribution does not the evaluation of γ and ϵ, hence the E-step does not change compared to the one in exercise 17.1.

As long as $\mathbf{A}$ and $\mathbf{B}$ are estimated independently, we are now focus on estimating $\mathbf{B} = (\pi, \mu, \Sigma)$ during M-step, the involved target function is:

$$\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\gamma_{i,t}(k)\log p(x_{i,t}|\mathbf{B})$$

Since the parameters are independent w.r.t k, we delve into a case where k is given. We also denote the iteration through $i = 1$ to N and $t = 1$ to T_i by $n = 1$ to $T = \sum_{i=1}^{N} T_i$, now the log-likelihood takes the form:

$$\sum_{n=1}^{T}\gamma_n(k)\log p(x_n|\pi_k, \mu_k, \Sigma_k)$$

It can be seen as a weighted form of log-likelihood for a mixture of Gaussian, assume the mixture contains C(it should be C_k, but this notation causes no contradiction as long as we take k for granted) Gaussians. We are to apply another EM procedure during the M-step for this HMM. Denote the latent variable corresponding to x_n by $h_{n,k}$. Estimate the distribution of $p(h_{n,k}|z_n, \pi_k, \mu_k, \Sigma_k)$ is tantamount to the E-step used in handling traditional mixture of Gaussians. Denote the expectation of $h_{n,k}$'s components by $\gamma'_{c,n}(k)$.

Now applying the M-step of mixture of Gaussians, recall that auxiliay takes the form:

$$\sum_{n=1}^{T} \gamma_n(k) \sum_{c=1}^{C} \gamma'_{c,n}(k) \left\{\log \pi_{k,c} + \log N(x_n|\mu_{k,c}, \Sigma_{k,c})\right\}$$

Hence this HMM reweighted a traditional mixture of Gaussians, with the weight changed from $\gamma'_{c,n}(k)$ into $\gamma_n(k) \cdot \gamma'_{c,n}(k)$. The rest estimation is trivially the application of M-step in mixture of Gaussians using new weights.

17.4 EM for HMMs with tied mixtures

Recall the conclusion from exercise 17.3, the last M-step inside M-step takes the form:

$$\sum_{k=1}^{K} \sum_{n=1}^{T} \sum_{c=1}^{C} \gamma_{c,n}(k) \left\{\log \pi_{k,c} + \log N(x_n|\mu_c, \Sigma_c)\right\}$$

Where we accordingly update the meaning of γ, and we also remove k from the footnotes of μ and Σ given the conditions in this exercise.

It is easy to notice that this target function again takes the form of M-step target for a traditional mixture of Gaussians. Taking independent k and update π_k gives the learning process of K mixing weights. Sum out k and C independent Gaussian parameters can be updated.

18 State space models

18.1 Derivation of EM for LG-SSM

We directly work on the auxiliary function:

$$
\begin{aligned}
Q(\theta, \theta^{old}) =& \mathbb{E}_{p(\mathbf{Z}|\mathbf{Y},\theta^{old})}[\log \prod_{n=1}^{N} p(z_{n,1:T_n}, y_{n,1:T_n}|\theta)] \\
=& \mathbb{E}[\sum_{n=1}^{N} \log p(z_{n,1}) \prod_{i=2}^{T_n} p(z_{n,i}|z_{n,i-1}) \prod_{i=1}^{T_n} p(y_{n,i}|z_{n,i})] \\
=& \mathbb{E}[\sum_{n=1}^{N} \log N(z_{n,1}|\mu_0, \Sigma_0) + \sum_{i=2}^{T_n} N(z_{n,i}|A_i z_{n,i-1} + B_i u_i, Q_i) \\
&+ \sum_{i=1}^{T_n} N(y_{n,i}|C_i z_{n,i} + D_i u_i, R_i)] \\
=& \mathbb{E}[N \log \frac{1}{|\Sigma_0|^{\frac{1}{2}}} + \left\{ -\frac{1}{2} \sum_{n=1}^{N} (z_{n,1} - \mu_0)^T \Sigma_0^{-1} (z_{n,1} - \mu_0) \right\} \\
&+ \sum_{i=2}^{T} N_i \log \frac{1}{|Q_i|^{\frac{1}{2}}} \\
&+ \left\{ -\frac{1}{2} \sum_{n=1}^{N_i} (z_{n,i} - A_i z_{n,i-1} - B_i u_i)^T Q_i^{-1} (z_{n,i} - A_i z_{n,i-1} - B_i u_i) \right\}] \\
&+ \sum_{i=2}^{T} N_i \log \frac{1}{|R_i|^{\frac{1}{2}}} \\
&+ \left\{ -\frac{1}{2} \sum_{n=1}^{N_i} (y_{n,i} - C_i z_{n,i} - D_i u_i)^T R_i^{-1} (y_{n,i} - C_i z_{n,i} - D_i u_i) \right\}]
\end{aligned}
$$

When exchanging the order of sum over data, we have $T = \max_n \{T_n\}$ and N_i denotes the number of data set with size no more than i.

To estimate μ_0, take the related terms:

$$\mathbb{E}[-\frac{1}{2} \sum_{n=1}^{N} (z_{n,1} - \mu_0) \Sigma_0^{-1} (z_{n,1} - \mu_0)]$$

Take derivative w.r.t μ_0:

$$\mathbb{E}[\sum_{n=1}^{N} -\frac{1}{2} \mu_0^T \Sigma_0^{-1} \mu_0 + z_{n,1} \Sigma_0^{-1} \mu_0]$$

Setting it to zero yields:

$$\mu_0 = \frac{1}{N}\mathbb{E}[z_{n,1}]$$

It is obvious that such estimation is similar to that for MVN with x_n replaced by $\mathbb{E}[z_{n,1}]$. This similarity works for other parameters as well. For example, estimate Σ_0 is tantamount to estimate the covariance of MVN with data terms replaced.

Such analysis works for Q_i and R_i as well. To estimate coefficient matrix, we consider A_i firstly. The related term is:

$$\mathbb{E}[\sum_{n=1}^{N_i}\{z_{n,i}^T A_i^T Q_i^{-1} A_i z_{n,i} - 2z_{n,i-1}^T A_i^T Q_i^{-1}(z_{n,i} - B_i u_i)\}]$$

Setting derivative to zero yields a solution similar to that for μ_0, the same analysis can be applied for B_i, C_i, D_i as well.

18.2 Seasonal LG-SSM model in standard form

From Fig.18.6(a), we have:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & \mathbf{0}_{S-1}^T \\ 0 & 1 & 0 & \mathbf{0}_{S-1}^T \\ 0 & 0 & 1 & \mathbf{0}_{S-1}^T \\ \mathbf{0}_{S-1} & \mathbf{0}_{S-1} & \mathbf{I} & \mathbf{0}_{S-1} \end{pmatrix}$$

$$\mathbf{Q} = \begin{pmatrix} Q_a & \mathbf{0}_{S+1}^T & & \\ 0 & Q_b & \mathbf{0}_S^T & \\ 0 & 0 & Q & \mathbf{0}_{S-1}^T \\ \mathbf{0}_{(S-1)*(S+2)} & & & \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & \mathbf{0}_{S-1}^T \end{pmatrix}$$

Where we use $\mathbf{0}_n$ to denote a colomn vector of 0 with length n, and $\mathbf{0}_{m*n}$ to denote a $m * n$ matrix of 0.

19 Undirected graphical models(Markov random fields)

19.1 Derivation of the log partition function

According to the definition:

$$Z(\theta) = \sum_{\mathbf{y}} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c)$$

It is straightforward to give:

$$\begin{aligned}
\frac{\partial \log Z(\theta)}{\partial \theta_{c'}} =& \frac{\partial}{\partial \theta_{c'}} \log \sum_{\mathbf{y}} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c) \\
=& \frac{1}{Z(\theta)} \sum_{\mathbf{y}} \frac{\partial}{\partial \theta_{c'}} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c) \\
=& \frac{1}{Z(\theta)} \sum_{\mathbf{y}} \prod_{c \in C, c \neq c'} \psi_c(\mathbf{y}_c|\theta_c) \frac{\partial}{\partial \theta_{c'}} \psi_{c'}(\mathbf{y}_{c'}|\theta_{c'}) \\
=& \frac{1}{Z(\theta)} \sum_{\mathbf{y}} \prod_{c \in C, c \neq c'} \psi_c(\mathbf{y}_c|\theta_c) \frac{\partial}{\partial \theta_{c'}} \exp\left\{\theta_{c'}^T \phi_{c'}(\mathbf{y}_{c'})\right\} \\
=& \frac{1}{Z(\theta)} \sum_{\mathbf{y}} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c) \phi_{c'}(\mathbf{y}_{c'}) \\
=& \sum_{\mathbf{y}} \phi_{c'}(\mathbf{y}_{c'}) \frac{1}{Z(\theta)} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta) \\
=& \sum_{\mathbf{y}} \phi_{c'}(\mathbf{y}_{c'}) p(\mathbf{y}|\theta) \\
=& \mathbb{E}[\phi_{c'}(\mathbf{y}_{c'})|\theta]
\end{aligned}$$

19.2 CI properties of Gaussian graphical models

Problem a:

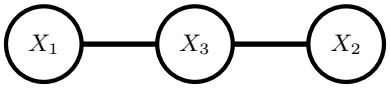
We have:

$$\Sigma = \begin{pmatrix} 0.75 & 0.5 & 0.25 \\ 0.5 & 1.0 & 0.5 \\ 0.25 & 0.5 & 0.75 \end{pmatrix}$$

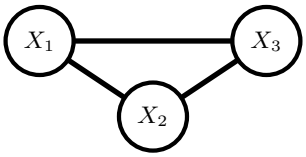
And:

$$\Lambda = \Sigma^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

Thus we have independency: $X_1 \perp X_2 | X_3$. This introduces a MRF like:



Problem b: The inverse of Σ contains no zero element, hence no conditional independency. Therefore there have to be edges between any two vertexes.



This model also cancels the marginal independency $X_1 \perp X_3$. But it is possible to model this set of properties by Bayesian network with two directed edges $X_1 \to X_2$ and $X_3 \to X_2$.

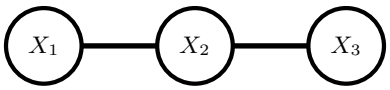
Problem c: Consider the terms inside the exponential:

$$-\frac{1}{2}\left\{x_1^2 + (x_2 - x_1)^2 + (x_3 - x_2^2)\right\}$$

It is easy to see the precision matrix and covariance matrix take:

$$\Lambda = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

Problem d: The only independency is $X_1 \perp X_3 | X_2$:



19.3 Independencies in Gaussian graphical models

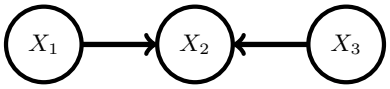
Problem a and b:

This PGM implies $X_1 \perp X_3|X_2$, hence we are looking for a precision matrix with $\Lambda_{1,3} = 0$, thus C and D meet the condition. On the other hand, $(A^{-1})_{1,3} = (B^{-1})_{1,3} = 0$. So A and B are candidates for covariance matrix.

Problem c and d:

This PGM tells that $X_1 \perp X_3$. Hence C and D can be covariance matrix, A and B can be precision matrix.

The only possible PGM is:



Problem e:

The answer can be derived from the conclusion of marginal Gaussian directly, A is true while B not.

19.4 Cost of training MRFs and CRFs

The answer are generally:

$$O(r(Nc+1))$$

and

$$O(r(Nc+N))$$

19.5 Full conditional in an Ising model

Straightforwardly(we have omitted θ from condition w.l.o.g):

$$\begin{aligned}
p(x_k = 1|\mathbf{x}_{-k}) =& \frac{p(x_k = 1, \mathbf{x}_{-k})}{p(\mathbf{x}_{-k})} \\
=& \frac{p(x_k = 1, \mathbf{x}_{-k})}{p(x_k = 0, \mathbf{x}_{-k}) + p(x_k = 1, \mathbf{x}_{-k})} \\
=& \frac{1}{1 + \frac{p(x_k=0, \mathbf{x}_{-k})}{p(x_k=1, \mathbf{x}_{-k})}} \\
=& \frac{1}{1 + \frac{\exp(h_k \cdot 0) \prod_{<k,i>} \exp(J_{k,i} \cdot 0)}{\exp(h_k \cdot 1) \prod_{<k,i>} \exp(J_{k,i} \cdot x_i)}} \\
=& \sigma(h_k + \sum_{i=1, i \neq k}^{n} J_{k,i} x_i)
\end{aligned}$$

When using denotation $x = \{0, 1\}$, the full conditional becomes:

$$p(x_k = 1|\mathbf{x}_{-k}) \sigma(2 \cdot (h_k + \sum_{i=1, i \neq k}^{n} J_{k,i} x_i))$$

20 Exact inference for graphical models

20.1 Variable elimination

Where tf is the figure?!

20.2 Gaussian times Gaussian is Gaussian

We have:

$$\begin{aligned}
&N(x|\mu_1, \lambda_1^{-1}) \times NN(x|\mu_2, \lambda_2^{-1}) \\
=&\frac{\sqrt{\lambda_1\lambda_2}}{2\pi} \exp\left\{-\frac{\lambda_1}{2}(x-\mu_1)^2 - \frac{\lambda_2}{2}(x-\mu_2)^2\right\} \\
=&\frac{\sqrt{\lambda_1\lambda_2}}{2\pi} \exp\left\{-\frac{\lambda_1+\lambda_2}{2}x^2 + (\lambda_1\mu_1+\lambda_2\mu_2)x - \frac{\lambda_1\mu_1^2+\lambda_2\mu_2^2}{2}\right\}
\end{aligned}$$

By completing the square:

$$\begin{aligned}
&\exp\left\{-\frac{\lambda_1+\lambda_2}{2}x^2 + (\lambda_1\mu_1+\lambda_2\mu_2)x - \frac{\lambda_1\mu_1^2+\lambda_2\mu_2^2}{2}\right\} \\
=&c \cdot \exp -\frac{\lambda}{2}(x-\mu)^2
\end{aligned}$$

Where:

$$\lambda = \lambda_1 + \lambda_2$$

$$\mu = \lambda^{-1}(\lambda_1\mu_1 + \lambda_2\mu_2)$$

The constant factor c can be obtained by computing the constant terms inside the exponential.

20.3 Message passing on a tree

Problem a:

It is easy to see after variable elimination:

$$p(X_2 = 50) = \sum_{G_1}\sum_{G_2} p(G_1)p(G_2|G_1)p(X_2 = 50|G_2)$$

$$p(G_1 = 1, X_2 = 50) = p(G_1)\sum_{G_2} p(G_2|G_1 = 1)p(X_2 = 50|G_2)$$

Thus:

$$p(G_1 = 1|X_2 = 50) = \frac{0.45 + 0.05 \cdot \exp(-5)}{0.5 + 0.5 \cdot \exp(-5)} \approx 0.9$$

Problem b(here X denotes X_2 or X_3):

$$\begin{aligned}
&p(G_1 = 1|X_2 = 50, X_3 = 50) \\
=&\frac{p(G_1 = 1, X_2 = 50, X_3 = 50)}{p(X_2 = 50, X_3 = 50)} \\
=&\frac{p(G_1 = 1)p(X_2|G_1 = 1)p(X_3|G_1 = 1)}{p(G_1 = 0)p(X_2|G_1 = 0)p(X_3|G_1 = 0) + p(G_1 = 1)p(X_2|G_1 = 1)p(X_3|G_1 = 1)} \\
=&\frac{p(X = 50|G_1 = 1)^2}{p(X = 50|G_1 = 0)^2 + p(X = 50|G_1 = 1)^2} \\
\approx&\frac{0.9^2}{0.1^2 + 0.9^2} \approx 0.99
\end{aligned}$$

Extra evidence makes the belief in $G_1 = 1$ firmer.

Problem c:

The answer to problem c is symmetric to that to problem b, $p(G_1 = 0|X_2 = 60, X_3 = 60) \approx 0.99$.

Problem d:

Using the same pattern of analysis from Problem b, we have:

$$\begin{aligned}
&p(G_1 = 1|X_2 = 50, X_3 = 60) \\
=&\frac{p(X = 50|G_1 = 1)p(X = 60|G_1 = 1)}{p(X = 50|G_1 = 0)p(X = 60|G_1 = 0) + p(X = 50|G_1 = 1)p(X = 60|G_1 = 1)}
\end{aligned}$$

Notice we have:

$$p(X = 50|G_1 = 1) = p(X = 60|G_1 = 0)$$

$$p(X = 50|G_1 = 0) = p(X = 60|G_1 = 1)$$

Hence:

$$P(G_1 = 1|X_2 = 50, X_3 = 60) = 0.5$$

In this case, X_2 and X_3 have equal strength as evidence and their effects achieve a balance so they provide not enough information to distort the prior knowledge.

20.4 Inference in 2D lattice MRFs

Please refer to PGM:principals and techniques 11.4.1.

21 Variational inference

21.1 Laplace approximation to $p(\mu, \log\sigma|D)$ for a univariate Gaussian

Laplace approximation equals representing $f(\mu, l) = \log p(\mu, l = \log\sigma|D)$ with second-order Taylor expansion. We have:

$$\begin{aligned}
\log p(\mu, l|D) &= \log p(\mu, l, D) - \log p(D) \\
&= \log p(\mu, l) + \log p(D|\mu, l) + c \\
&= \log p(D|\mu, l) + c \\
&= \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \mu)^2\right\} + c \\
&= -N\log\sigma + \sum_{n=1}^{N} -\frac{1}{2\sigma^2}(y_n - \mu)^2 + c \\
&= -N\cdot l + \frac{1}{2}\frac{1}{\exp\{2\cdot l\}} \sum_{n=1}^{N}(y_n - \mu)^2 + c
\end{aligned}$$

Thus we derive:

$$\begin{aligned}
\frac{\partial \log p(\mu, l|D)}{\partial \mu} &= \frac{1}{2}\frac{1}{\exp\{2\cdot l\}} \sum_{n=1}^{N} 2\cdot(y_n - \mu) \\
&= \frac{N}{\sigma^2}\cdot(\bar{y} - \mu) \\
\frac{\partial \log p(\mu, l|D)}{\partial l} &= -N + \frac{1}{2}\sum_{n=1}^{N}(y_n - \mu)^2 \cdot (-2) \cdot \frac{1}{\exp\{2\cdot l\}} \\
&= -N + \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n - \mu)^2 \\
\frac{\partial^2 \log p(\mu, l|D)}{\partial \mu^2} &= -\frac{N}{\sigma^2} \\
\frac{\partial^2 \log p(\mu, l|D)}{\partial l^2} &= -\frac{2}{\sigma^2}\sum_{n=1}^{N}(y_n - \mu)^2 \\
\frac{\partial^2 \log p(\mu, l|D)}{\partial \mu \partial l} &= N\cdot(\bar{y} - \mu)\cdot(-2)\cdot\frac{1}{\sigma^2}
\end{aligned}$$

For approximation, $p(\mu, l) \approx N(\mu, \Sigma)$ with:

$$\Sigma = \begin{pmatrix} \frac{\partial^2 \log p(\mu,l|D)}{\partial \mu^2} & \frac{\partial^2 \log p(\mu,l|D)}{\partial l^2} \\ \frac{\partial^2 \log p(\mu,l|D)}{\partial l^2} & \frac{\partial^2 \log p(\mu,l|D)}{\partial \mu \partial l} \end{pmatrix}^{-1}$$

$$\mu = \Sigma \begin{pmatrix} \frac{\partial \log p(\mu,l|D)}{\partial \mu} \\ \frac{\partial \log p(\mu,l|D)}{\partial l} \end{pmatrix}$$

21.2 Laplace approximation to normal-gamma

This is the same with exercise 21.1 when the prior is uniformative. We formally substitute:

$$\begin{aligned} \sum_{n=1}^{N}(y_n - \mu)^2 =& \sum_{n=1}^{N}((y_n - \bar{y}) - (\mu - \bar{y}))^2 \\ =& \sum_{n=1}^{N}(y_n - \bar{y})^2 + \sum_{n=1}^{N}(\mu - \bar{y})^2 + 2(\mu - \bar{y}) \cdot \sum_{n=1}^{N}(y_n - \bar{y}) \\ =& Ns^2 + N(\mu - \bar{y})^2 \end{aligned}$$

Where $s^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})^2$

Conclusions in all problems a, b and c are included in the previous solution.

21.3 Variational lower bound for VB for univariate Gaussian

What left in section 21.5.1.6 is the derivation for 21.86 to 21.91. We omit the derivation for entropy for Gaussian and moments, which can be found in any information theory textbook. Now we derive the $\mathbb{E}[\ln x|x \sim Ga(a,b)]$, which can therefore yields to the entropy for a Gamma distribution.

We know that Gamma distribution is an exponential family distribution:

$$\begin{aligned} Ga(x|a,b) =& \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-b \cdot x\} \\ \propto& \exp\{-b \cdot x + (a-1)\ln x\} \\ =& \exp\left\{\phi(x)^T \theta\right\} \end{aligned}$$

The sufficient statistics is $\phi(x) = (x, \ln x)^T$ and natural parameter is given by $\theta = (-b, a-1)^T$. Thus Gamma distribution can be seen as the maximum entropy distribution under constraints on x and $\ln x$.

The culumant function is given by:

$$\begin{aligned} A(\theta) &= \log Z(\theta) \\ &= \log \frac{\Gamma(a)}{b^a} \\ &= \log \Gamma(a) - a \log b \end{aligned}$$

The expectation of sufficient statistics is given by the derivative of cumulant function, therefore:

$$\mathbb{E}[\ln x] = \frac{\partial A}{\partial (a-1)} = \frac{\Gamma'(a)}{\Gamma(a)} - \log b$$

According to defintion $\psi(a) = \frac{\Gamma'(a)}{\Gamma(a)}$:

$$\mathbb{E}[\ln x] = \psi(a) - \log b$$

The rest derivations are completed or trivial.

21.4 Variational lower bound for VB for GMMs

The lower bound is given by:

$$\begin{aligned} \mathbb{E}_q[\log \frac{p(\theta, D)}{q(\theta)}] =& \mathbb{E}_q[\log p(\theta, D)] - \mathbb{E}_q[q(\theta)] \\ =& \mathbb{E}_q[\log p(D|\theta)] + \mathbb{E}_q[\log p(\theta)] + \mathbb{E}_q[\log q(\theta)] \\ =& \mathbb{E}[\log p(\mathbf{x}|\mathbf{z}, \mu, \Lambda, \pi)] + \mathbb{E}[\log p(\mathbf{z}, \mu, \Lambda, \pi)] \\ & - \mathbb{E}[\log q(\mathbf{z}, \mu, \Lambda, \pi)] \\ =& \mathbb{E}[\log p(\mathbf{x}|\mathbf{z}, \mu, \Lambda, \pi)] + \mathbb{E}[\log p(\mathbf{z}|\pi)] + \mathbb{E}[\log p(\pi)] + \mathbb{E}[\log p(\mu, \Lambda)] \\ & + \mathbb{E}[\log q(\mathbf{z})] + \mathbb{E}[\log q(\pi)] + \mathbb{E}[\log q(\mu, \Lambda)] \end{aligned}$$

We are now showing 21.209 to 21.215.

For 21.209:

$$\begin{aligned} \mathbb{E}[\log p(\mathbf{x}|\mathbf{z}, \mu, \Lambda)] =& \mathbb{E}_{q(\mathbf{z})q(\mu,\Lambda)}[\log p(\mathbf{x}|\mathbf{z}, \mu, \Lambda)] \\ =& \sum_n \sum_k \mathbb{E}_{q(\mathbf{z})q(\mu,\Lambda)}[-\frac{D}{2} \log 2\pi + \frac{1}{2} \log |\Lambda_k| - \frac{1}{2}(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] \end{aligned}$$

Using 21.132 and converting summing by average $\bar{x}_k$ yields to solution.
For 21.210:

$$\begin{aligned}
\mathbb{E}[\log p(\mathbf{z}|\pi)] =& \mathbb{E}_{q(\mathbf{z})q(\pi)}[\log p(\mathbf{z}|\pi)] \\
=& \mathbb{E}_{q(\mathbf{z})q(\pi)}[\log \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}] \\
=& \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{q(\mathbf{z})q(\pi)}[z_{nk} \log \pi_k] \\
=& \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{q(\mathbf{z})}[z_{nk}]\mathbb{E}_{q(\pi)}[\log \pi_k] \\
=& \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \log \bar{\pi}_k
\end{aligned}$$

For 21.211:

$$\begin{aligned}
\mathbb{E}[\log p(\pi)] =& \mathbb{E}_{q(\pi)}[\log p(\pi)] \\
=& \mathbb{E}_{q(\pi)}[\log(C \cdot \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1})] \\
=& \ln C + (\alpha_0 - 1) \sum_{k=1}^{K} \log \bar{\pi}_k
\end{aligned}$$

For 21.212:

$$\begin{aligned}
\mathbb{E}[\log p(\mu, \Lambda)] =& \mathbb{E}_{q(\mu,\Lambda)}[\log p(\mu, \Lambda)] \\
=& \mathbb{E}_{q(\mu,\Lambda)}[\log \prod_{k=1}^{K} Wi(\Lambda_k|L_0, v_0) \cdot N(\mu_k|m_0, (\beta_0\Lambda_k)^{-1}] \\
=& \sum_{k=1}^{K} \mathbb{E}_{q(\mu,\Lambda)}[\log C + \frac{1}{2}(v_0 - D - 1) \log |\Lambda_k| - \frac{1}{2} tr\left\{\Lambda_k L_0^{-1}\right\} \\
& - \frac{D}{2} \log 2\pi - \frac{1}{2} \log |\beta_0\Lambda_k| - \frac{1}{2}(\mu_k - m_0)^T(\beta_0\Lambda_k)(\mu_k - m_0)]
\end{aligned}$$

Using 21.131 to expand the expected value of the quadratic form and using the fact that the mean of a Wi distribution is $v_k L_k$ and we are done.

For 21.213:

$$\begin{aligned}\mathbb{E}[\log q(\mathbf{z})] =& \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] \\ =& \mathbb{E}_{q(\mathbf{z})}[\sum_i \sum_k z_{ik} \log r_{ik}] \\ =& \sum_i \sum_k \mathbb{E}_{q(\mathbf{z})}[z_{ik}] \log r_{ik} \\ =& \sum_i \sum_k r_{ik} \log r_{ik}\end{aligned}$$

For 21.214:

$$\begin{aligned}\mathbb{E}[\log q(\pi)] =& \mathbb{E}_{q(\pi)}[\log q(\pi)] \\ =& \mathbb{E}_{q(\pi)}[\log C + \sum_{k=1}^{K} (\alpha_k - 1) \log \pi_k] \\ =& \log C + \sum_k (\alpha_k - 1) \log \bar{\pi}_k\end{aligned}$$

For 21.215:

$$\begin{aligned}\mathbb{E}[\log q(\mu, \Lambda)] =& \mathbb{E}_{q(\mu,\Lambda)}[\log q(\mu, \Lambda)] \\ =& \sum_k \mathbb{E}_{q(\mu,\Lambda)}[\log q(\Lambda_k) - \frac{D}{2} \log 2\pi + \frac{1}{2} \log |\beta_k \Lambda_k| \\ & - \frac{1}{2}(\mu_k - m_k)^T (\beta_k \Lambda_k)(\mu_k - m_k)]\end{aligned}$$

Using 21.132 to expand the quadratic form to give $\mathbb{E}[(\mu_k - m_k)^T (\beta_k \Lambda_k)(\mu_k - m_k)] = D$

21.5 Derivation of $\mathbb{E}[\log \pi_k]$

under a Dirichlet distribution Dirichlet distribution is an exponential family distribution, we have:

$$\phi(\pi) = (\log \pi_1, \log \pi_2, ... \log \pi_K)$$

$$\theta = \alpha$$

The cumulant function is:

$$A(\alpha) = \log B(\alpha) = \sum_{i=1}^{K} \log \Gamma(\alpha_i) - \log \Gamma(\sum_{i=1}^{K} \alpha_i)$$

And:

$$\mathbb{E}[\log \pi_k] = \frac{\partial A(\alpha)}{\partial \alpha_k} = \frac{\Gamma'(\alpha_k)}{\Gamma(\alpha_k)} - \frac{\Gamma'(\sum_{i=1}^K \alpha_k)}{\Gamma(\sum_{i=1}^K \alpha_k)} = \psi(\alpha_k) - \psi(\sum_{i=1}^K \alpha_i)$$

Take exponential on both sides:

$$\exp(\mathbb{E}[\log \pi_k]) = \exp(\psi(\alpha_k) - \psi(\sum_{i=1}^K \alpha_k)) = \frac{\exp(\alpha_k)}{\exp(\sum_{i=1}^K \alpha_i)}$$

21.6 Alternative derivation of the mean field updates for the Ising model

This is no different than applying the procedure in section 21.3.1 before derivating updates, hence omitted.

21.7 Forwards vs reverse KL divergence

We have:

$$\begin{aligned}
KL(p(x,y)||q(x,y)) =& \mathbb{E}_{p(x,y)}[\log \frac{p(x,y)}{q(x,y)}] \\
=& \sum_{x,y} p(x,y) \log p(x,y) - \sum_{x,y} p(x,y) \log q(x) - \sum_{x,y} p(x,y) \log q(y) \\
=& \sum_{x,y} p(x,y) \log p(x,y) - \sum_{x} (\sum_{y} p(x,y)) \log q(x) - \sum y (\sum_{x} p(x,y)) \log q(q) \\
=& H(p(x,y)) - H(p(x)) - H(p(y)) + KL(p(x)||q(x)) + KL(p(y)||q(y)) \\
=& constant + KL(p(x)||q(x)) + KL(p(y)||q(y))
\end{aligned}$$

Thus the optimal approximation is $q(x) = p(x)$ and $q(y) = p(y)$.

We skip the practical part.

21.8 Derivation of the structured mean field updates for FHMM

According to the conclusion from mean-field varitional methods, we have:

$$E(\mathbf{x}_m) = \mathbb{E}_{q/m}[E(\bar{p}(\mathbf{x}_m))]$$

Thus:

$$-\sum_{t=1}^{T}\sum_{k=1}^{K} x_{t,m,k}\tilde{\epsilon}_{t,m,k} = \frac{1}{2}\mathbb{E}[\sum_{t=1}^{T}(\mathbf{y}_t - \sum_{l\neq m}^{M} W_l\mathbf{x}_{t,m})^T\Sigma^{-1}(\mathbf{y}_t - \sum_{l\neq m}^{M} W_l\mathbf{x}_{t,m})] + C$$

Comparing the coefficient of $x_{t,m,k}$ (i.e. setting $x_{t,m,k}$ to 1) ends in:

$$\tilde{\epsilon}_{t,m,k} = W_m^T\Sigma^{-1}(\mathbf{y}_t - \sum_{l\neq m} W_l\mathbb{E}[\mathbf{x}_{t,l}]) - \frac{1}{2}(W_m^T\Sigma^{-1}W_m)_{k,k}$$

Write into matrix form yields to 21.62.

21.9 Variational EM for binary FA with sigmoid link

Refer to "Probabilistic Visualisation of High-Dimensional Binary Data, Tipping, 1998".

21.10 VB for binary FA with probit link

The major difference in using probit link is the uncontinuous likelihood caused by $p(y_i = 1|z_i) = \mathbb{I}(z_i > 0)$. In the context of hiding $\mathbf{X}$, we assume Gaussian prior on $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{Z}$. The approximation takes the form:

$$q(\mathbf{X}, \mathbf{Z}, \mathbf{W}) = \prod_{l=1}^{L} q(\mathbf{w}_l)\prod_{i=1}^{N} q(\mathbf{x}_i)q(z_i)$$

It is a mean-field approximation, hence in an algorithm similari to EM, we are to update the distribution of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ stepwise.

For variable $\mathbf{X}$, we have:

$$\begin{aligned}\log q(\mathbf{x}_i) =& \mathbb{E}_{q(\mathbf{z}_i)q(\mathbf{w})}[\log p(\mathbf{x}_i, \mathbf{w}, z_i, y_i)] \\ =& \mathbb{E}_{q(\mathbf{z}_i)q(\mathbf{w})}[\log p(\mathbf{x}_i) + \log p(\mathbf{w}) + \log p(z_i|\mathbf{w}_i, \mathbf{w}) + \log p(y_i|z_i)]\end{aligned}$$

Given the likelihood form, for i corresponding to $y_i = 1$, $q(z_i)$ have to be a truncated one, i.e. we only consider the expectations in the form $\mathbb{E}[z|z > \mu]$ and $\mathbb{E}[z^2|z > \mu]$.

$\log q(\mathbf{x}_i) = -\frac{1}{2}\mathbf{x}_i^T\Lambda_1\mathbf{x}_i - \frac{1}{2}\mathbb{E}[z^2] - \frac{1}{2}\mathbf{x}_i^T\mathbb{E}[\mathbf{w}\mathbf{w}^T]\mathbf{x}_i + \mathbb{E}[z]\mathbb{E}[\mathbf{w}]^T\mathbf{x}_i$

Where Λ_1 is the covariance of $\mathbf{x}_i$'s prior distribution, $\mathbb{E}[\mathbf{w}\mathbf{w}^T]$ can be calculated given the Gaussian form of $q(\mathbf{w})$, and truncated expectations $\mathbb{E}[z]$

and $\mathbb{E}[z^2]$ can be obtained from solutions to exercise 11.15. It is obvious that $q(\mathbf{x}_i)$ is a Gaussian.

The update for $\mathbf{w}$ is similar to that for $\mathbf{x}_i$ as long as they play symmetric roles in likelihood. The only difference is we have to sum over i when updating $\mathbf{w}$.

At last we update z_i:

$$\log q(z_i) = \mathbb{E}_{q(\mathbf{x}_i)q(\mathbf{w})}[\log p(z_i|\mathbf{x}_i, \mathbf{w}) + \log p(y_i|z_i)]$$

Inside the expectation we have:

$$-\frac{1}{2}z_i^2 + \mathbb{E}[\mathbf{w}]^T\mathbb{E}[\mathbf{x}]z_i + c$$

Therefore $q(z_i)$ again takes a Gaussian form.