More PRML Errata

Yousuke Takada
yousuketakada@gmail.com

June 14, 2017

Preface

This report communicates some more errata for *Pattern Recognition and Machine Learning* or PRML by Bishop (2006) that are not listed in the official errata document (Svensén and Bishop, 2011)[1] at the time of this writing. When specifying the location of an error ("Paragraph 2, Line –1" or the like), I follow the notational conventions adopted by Svensén and Bishop (2011). As the official errata document "is intended to be complete," this report also tries to correct even trivial typographical errors as well.

PRML is arguably such a great textbook in the field of machine learning that it is extremely helpful and easier to understand than any other similar account. That said, there are a few subtleties that some readers (including I) might have hard time to appreciate. In hopes to help such readers get out of struggle or become more confident about important concepts, I have also included in this report some comments and suggestions for improving the readability to which I would have liked to refer when I first read PRML.

It should also be noted that the readers of the Japanese edition of PRML will find its support page (in Japanese) useful. Along with other information, it lists errata specific to the Japanese edition as well as some additional errata for the English edition, which have also been included in this report for the reader's convenience.

I welcome all comments and suggestions regarding this report; please send me any such feedback via email or, preferably, by creating an "issue" or a "pull request" at the following GitHub repository

https://github.com/yousuketakada/prml_errata

where you can find the source code of this report as well as other supporting material.

Acknowledgements

I would like to thank those who have kindly informed me of yet more errata and clarifications incorporated in this report. In particular, I am grateful to Christopher Sahnwaldt and Mark-Jan Nederhof for their invaluable discussions.

[1] The last line but one of the bibliographic information page of the copy of PRML I have reads "9 8 7 (corrected at 6th printing 2007)." So I refer to Version 2 of the errata document (Svensén and Bishop, 2011).

Corrections

Page xi

Paragraph –2, Line 1: $|f(x)/g(x)|$ should read $|g(x)/f(x)|$ (with the functions swapped). Moreover, the limit we take is *not* necessarily the one specified in the text, i.e., $x \to \infty$, but is often implied by the context as follows.

Big O notation The big O notation $g(x) = O\left(f(x)\right)$ generally denotes that $|g(x)/f(x)|$ is bounded as $x \to c$ where, if c is not given explicitly, $c = 0$ for a Taylor series such as (2.299) or (D.1); or $c = \infty$ for an asymptotic series such as (10.241) or for computational complexity (see, e.g., Section 5.2.3), for example. See Olver et al. (2016) for other asymptotic and order notations.

Page 5

Equation (1.1): The lower ellipsis (. . .) should be centered (· · ·).[2] Specifically, (1.1) should read

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j. \tag{1}$$

Page 10

The text after (1.4): The lower ellipsis (. . .) should be centered (· · ·).

Page 46

Equation (1.85): A period (.) should terminate (1.85).

Page 51

Equation (1.98): Following the notation (1.93), we should write the left hand side of (1.98) as $\mathrm{H}[X]$ instead of $\mathrm{H}[p]$. As suggested in the first paragraph of Appendix D on Page 703, if we regard the entropy $\mathrm{H}[\cdot]$ as a functional, we see that "the entropy could equally well have been written as $\mathrm{H}[p]$." However, it is probably better to maintain the notational consistency here.

Page 56

Equation (1.116): In general, we cannot interpret λ_i in *Jensen's inequality* (1.115) as the probability distribution over a discrete random variable x such that $\lambda_i \equiv p(x = x_i)$ because, since (1.115) holds for any $\{x_i\}$, we can take, say, $x_i = x_j$ and $\lambda_i \neq \lambda_j$ where $i \neq j$, assigning different probabilities for the same value of x. Actually, (1.116) is a special case of (1.115). An equivalent of (1.115) in terms of random variables can be derived as follows.

[2] The LaTeX command `\cdots` or, with the `amsmath` or `mathtools` package, `\dots` will do.

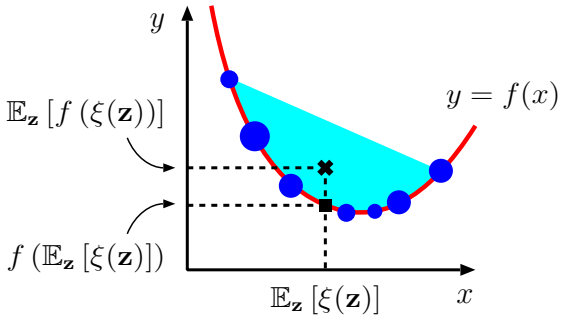

Figure 1 A physical "proof" of Jensen's inequality (MacKay, 2003). Let us suppose that we have a set of point masses $m_i = p(\mathbf{z} = \mathbf{z}_i)$, denoted by filled blue circles (●) with areas proportional to m_i, and place them at the corresponding locations $(x, y) = (\xi(\mathbf{z}_i), f(\xi(\mathbf{z}_i)))$ on a convex curve $y = f(x)$. The center of gravity of those masses, which is $(\mathbb{E}_{\mathbf{z}}\left[\xi(\mathbf{z})\right], \mathbb{E}_{\mathbf{z}}\left[f\left(\xi(\mathbf{z})\right)\right])$, denoted by a cross sign (×), must lie above the convex curve and thus right above the point $(\mathbb{E}_{\mathbf{z}}\left[\xi(\mathbf{z})\right], f\left(\mathbb{E}_{\mathbf{z}}\left[\xi(\mathbf{z})\right]\right))$ on the curve, denoted by a filled square (■), showing Jensen's inequality (2). One can also see that, if $f(\cdot)$ is strictly convex, the equality in (2) implies that $\xi(\cdot)$ is essentially constant (it is trivial to show that the converse is true).

Jensen's inequality in terms of random variables In order to interpret (1.115) probabilistically, we instead introduce another set of underlying random variables $\mathbf{z}$ such that $\lambda_i \equiv p(\mathbf{z} = \mathbf{z}_i)$ and a function $\xi(\cdot)$ such that $x_i \equiv \xi(\mathbf{z}_i)$, giving a result slightly more general than (1.116)

$$f\left(\mathbb{E}_{\mathbf{z}}\left[\xi(\mathbf{z})\right]\right) \leqslant \mathbb{E}_{\mathbf{z}}\left[f\left(\xi(\mathbf{z})\right)\right] \tag{2}$$

where $f(\cdot)$ is a convex function but $\xi(\cdot)$ can be any. Moreover, if $f(\cdot)$ is strictly convex, the equality in (2) holds if and only if $\xi(\cdot)$ is *essentially constant*, meaning that there exists some constant ξ_0 and $\xi(\mathbf{z}) = \xi_0$ for almost[3] all $\mathbf{z}$ such that $p(\mathbf{z}) > 0$ (in other words, if we regard $\xi = \xi(\mathbf{z})$ as a random variable, we have $\xi = \xi_0$ almost surely, i.e., with probability one), in which case we have $\mathbb{E}_{\mathbf{z}}\left[\xi(\mathbf{z})\right] = \xi_0$ and the both sides of (2) equal $f(\xi_0)$. See Figure 1 for an intuitive, physical "proof" of the inequality (2).

Since the random variables $\mathbf{z}$ as well as their probability $p(\mathbf{z})$ can be chosen arbitrarily, it makes sense to write $\mathbf{z}$ implicit in (2), giving a simpler form of Jensen's inequality

$$f\left(\mathbb{E}\left[\xi\right]\right) \leqslant \mathbb{E}\left[f\left(\xi\right)\right]. \tag{3}$$

For continuous random variables, we have

$$f\left(\int \xi(\mathbf{x}) p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \leqslant \int f\left(\xi(\mathbf{x})\right) p(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{4}$$

where we have used $\mathbf{x}$ to denote the underlying random variables for which we take the expectations. By making use of (4), one can show that the *Kullback-Leibler divergence* $\mathrm{KL}\left(p\|q\right)$ given by (1.113) satisfies *Gibbs's inequality*

$$\mathrm{KL}\left(p\|q\right) \geqslant 0 \tag{5}$$

with equality if and only if $p(\mathbf{x}) = q(\mathbf{x})$ almost everywhere. See the following erratum for more details.

[3]Here, the term "almost" means that there may be some exceptions but they can occur only with probability zero so that we can safely ignore them. As in PRML, we omit the term "almost" for brevity in the rest of this report.

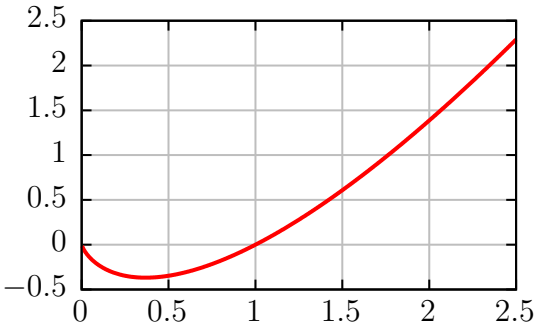

Figure 2 Plot of $f(x) = x \ln x$. The function $f(x)$ is a strictly convex function defined over $[0, \infty)$ where we have defined $f(0) = 0 \ln 0 = 0$. The curve $y = f(x)$ takes the minimum at $(x, y) = (e^{-1}, -e^{-1})$. The roots (the values of x such that $f(x) = 0$) are $x = 0$ and $x = 1$.

Page 56

Equation (1.118): There are some difficulties in the derivation (1.118) of Gibbs's inequality (5). First, the quantity $\xi(\mathbf{x}) = q(\mathbf{x})/p(\mathbf{x})$ is undefined for $\mathbf{x}$ such that $p(\mathbf{x}) = 0$. Second, the convex function $f(\xi) = -\ln \xi$ is undefined for $\xi = 0$, which occurs where $q(\mathbf{x}) = 0$ and $p(\mathbf{x}) > 0$. In order to avoid these difficulties, we should take a different approach (MacKay, 2003; Kullback and Leibler, 1951) in which we make use of Jensen's inequality (4) with respect to $q(\mathbf{x})$ where we identify $f(\xi) = \xi \ln \xi$ (see Figure 2) and $\xi(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$. Note that we can safely proceed with this approach because we can assume $q(\mathbf{x}) > 0$ without loss of generality as we shall see in the following.

A proof of Gibbs's inequality Let us first examine the behavior of the integrand of the Kullback-Leibler divergence $\mathrm{KL}\,(p\|q)$

$$p(\mathbf{x}) \ln p(\mathbf{x}) - p(\mathbf{x}) \ln q(\mathbf{x}) \tag{6}$$

where $q(\mathbf{x})$ or $p(\mathbf{x})$ vanishes. We notice that, if $q(\mathbf{x}) \to 0$ for $\mathbf{x}$ such that $p(\mathbf{x}) > 0$, the integrand (6) diverges so that $\mathrm{KL}\,(p\|q) \to \infty$. On the other hand, the integrand (6) always vanishes for $\mathbf{x}$ such that $p(\mathbf{x}) = 0$ regardless of the values of $q(\mathbf{x})$.[4] Therefore, in order for $\mathrm{KL}\,(p\|q)$ to be well-defined, one must have $p(\mathbf{x}) = 0$ wherever $q(\mathbf{x}) = 0$; this property is called *zero-forcing* (see also Section 10.1.2).

Assuming the zero-forcing condition, that is, $p(\mathbf{x}) = 0$ for all $\mathbf{x}$ such that $q(\mathbf{x}) = 0$, we can consider the integration in $\mathrm{KL}\,(p\|q)$ only over $\Omega = \{\mathbf{x} \mid q(\mathbf{x}) > 0\}$. Identifying $f(\xi) = \xi \ln \xi$ and $\xi(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$, we have

$$\begin{aligned}
\mathrm{KL}\,(p\|q) &= \int_\Omega q(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} \ln \left\{ \frac{p(\mathbf{x})}{q(\mathbf{x})} \right\} \mathrm{d}\mathbf{x} && (7)\\
&= \int_\Omega q(\mathbf{x}) f\left(\xi(\mathbf{x})\right) \mathrm{d}\mathbf{x} && (8)\\
&\geqslant f\left(\int_\Omega q(\mathbf{x}) \xi(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right) && (9)\\
&= f\left(\int_\Omega p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right) && (10)\\
&= f(1) = 0 && (11)
\end{aligned}$$

[4] Recall that we have defined $0 \ln 0 = 0$ so that the entropy $\mathrm{H}[x]$ (1.93) is well-defined.

where we have used (4) with respect to $q(\mathbf{x})$ (instead of $p(\mathbf{x})$). Note that, since $q(\mathbf{x}) > 0$ over Ω, we see that $\xi(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x}) \geqslant 0$ is well-defined over Ω and so is $f\left(\xi(\mathbf{x})\right) = \xi(\mathbf{x}) \ln \xi(\mathbf{x})$. Since $f(\xi)$ is strictly convex, the equality $\mathrm{KL}\left(p \| q\right) = 0$ holds if and only if $\xi(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$ is constant over Ω, which, together with the zero-forcing condition, yields the equality condition that $p(\mathbf{x}) = q(\mathbf{x})$ for all $\mathbf{x}$.

Page 62

Exercise 1.18, the text after (1.142): "Gamma" should read "gamma" (without capitalization).

Page 70

Paragraph –1, Line –1: The lower ellipsis (. . .) should be centered ($\cdots$).

Page 78

The caption of Figure 2.5: "$\{\alpha_k\} = 0.1$" should read "$\alpha_k = 0.1$ for all k" and so on.

Page 80

Equation (2.52): We usually take eigenvectors $\mathbf{u}_i$ to be the columns of $\mathbf{U}$ as in (C.37). If we follow this convention, (2.52) and the following text should read

$$\mathbf{y} = \mathbf{U}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) \tag{12}$$

where $\mathbf{U}$ is a matrix whose columns are given by $\mathbf{u}_i$ so that $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_D)$. From (2.46) it follows that $\mathbf{U}$ is an *orthogonal* matrix, i.e., it satisfies $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and hence also $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

Page 81

Equations (2.53) and (2.54): If we write the change of variable from $\mathbf{x}$ to $\mathbf{y}$ as (12) instead of (2.52), the Jacobian matrix $\mathbf{J} = (J_{ij})$ is simply given by $\mathbf{U}$. Equation (2.53) should read

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ij} \tag{13}$$

where U_{ij} is the (ij)-th element of $\mathbf{U}$. The square of the determinant of the Jacobian matrix (2.54) can then be evaluated as

$$|\mathbf{J}|^2 = |\mathbf{U}|^2 = \left|\mathbf{U}^{\mathrm{T}}\right| |\mathbf{U}| = \left|\mathbf{U}^{\mathrm{T}}\mathbf{U}\right| = |\mathbf{I}| = 1. \tag{14}$$

Page 81

The text after (2.54): Since the Jacobian matrix $\mathbf{J}$ is only assumed to be orthogonal here, the determinant of $\mathbf{J}$ can be either positive or negative so that we should write $|\mathbf{J}| = \pm 1$ instead of $|\mathbf{J}| = 1$.

Page 82

Equation (2.56): We should take the absolute value of the determinant for the same reason given above so that the factor $|\mathbf{J}|$ should read $|\det(\mathbf{J})|$. Note however that it is not recommended to write $||\mathbf{J}||$ to mean $|\det(\mathbf{J})|$ because $||\mathbf{J}||$ is confusingly similar to the matrix norm $\|\mathbf{J}\|$, which usually refers to the largest singular value of $\mathbf{J}$ (Golub and Van Loan, 2013). This notational inconsistency is caused by the abuse of the notation $|\cdot|$ for both the absolute value and the matrix determinant; if we always use $\det(\cdot)$ for the determinant, confusion will not arise and the notation be consistent.

Notation for absolute determinant An alternative solution to the problem of notational inconsistency mentioned above would be to explicitly define $|\mathbf{A}|$ as the absolute value of the determinant of a square matrix $\mathbf{A}$, i.e.,

$$|\mathbf{A}| \equiv |\det(\mathbf{A})| \tag{15}$$

so that we have $|\mathbf{J}| = 1$ and (2.56) holds as is. Note also that this notation (15) is mostly consistent in other part of PRML because we have $|\mathbf{A}| = \det(\mathbf{A})$ for any positive-semidefinite matrix $\mathbf{A} \succeq 0$ (see Appendix C) and most matrices for which we take determinants are in fact positive definite.[5] Such positive-definite matrices include the covariance $\boldsymbol{\Sigma}$ or the precision $\boldsymbol{\Lambda}$ of the multivariate Gaussian distribution and the scale matrix $\mathbf{W}$ of the Wishart distribution (see Appendix B).

Page 90

Paragraph –1, Line 2: The partitioned vector should read $\mathbf{x} = \left(\mathbf{x}_a^{\mathrm{T}}, \mathbf{x}_b^{\mathrm{T}}\right)^{\mathrm{T}}$.

Pages 93 and 94

Equations (2.121) and (2.122): We obtain the maximum likelihood solutions $\boldsymbol{\mu}_{\mathrm{ML}}$ and $\boldsymbol{\Sigma}_{\mathrm{ML}}$ for the Gaussian by setting the derivatives of the log likelihood function $\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given by (2.118) with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ equal to zero, which, however, only implies that $\boldsymbol{\mu}_{\mathrm{ML}}$ and $\boldsymbol{\Sigma}_{\mathrm{ML}}$ are stationary points. We should also show that $\boldsymbol{\mu}_{\mathrm{ML}}$ and $\boldsymbol{\Sigma}_{\mathrm{ML}}$ indeed maximize the likelihood as discussed in the following.

Maximum likelihood for Gaussian Let us first maximize the likelihood function with respect to the mean $\boldsymbol{\mu}$. This can be easily done by noting that the log likelihood (2.118) is quadratic in $\boldsymbol{\mu}$ so that

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathrm{ML}}) + \text{const} \tag{16}$$

where $\boldsymbol{\mu}_{\mathrm{ML}}$ is given by (2.121) and the terms independent of $\boldsymbol{\mu}$ have been absorbed into "const." Since the covariance $\boldsymbol{\Sigma}$ is positive definite and so is its inverse $\boldsymbol{\Sigma}^{-1}$, we see that the log likelihood (16) is concave with respect to $\boldsymbol{\mu}$ and that $\boldsymbol{\mu}_{\mathrm{ML}}$ indeed maximizes the likelihood.

[5] In this report, we assume as customary that the concept of positive/negative (semi)definiteness is restricted to symmetric matrices. For example, when we say "$\mathbf{A}$ is positive definite" or $\mathbf{A} \succ 0$, we implicitly assume that $\mathbf{A}$ is also symmetric so that $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$, though we still sometimes say "$\mathbf{A}$ is symmetric positive definite" to avoid confusion.

Next, we consider maximization with respect to the covariance $\boldsymbol{\Sigma}$. The maximum likelihood solution $\boldsymbol{\Sigma}_{\mathrm{ML}}$ given by (2.122) can be obtained by solving

$$\nabla_{\boldsymbol{\Sigma}} \ln p\left(\mathbf{X}|\boldsymbol{\mu}_{\mathrm{ML}}, \boldsymbol{\Sigma}\right) = \mathbf{O} \tag{17}$$

where $\nabla_{\mathbf{A}}$ is the gradient operator with respect to a matrix $\mathbf{A}$ defined by (190) and $\mathbf{O}$ is a zero matrix. Making use of the eigenvalue expansion (2.48) of $\boldsymbol{\Sigma}$, we can write the log likelihood (2.118) in terms of the eigenvalues $\{\lambda_i\}$ so that

$$\ln p\left(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = -\frac{N}{2} \sum_{i=1}^{D} \left\{ \ln \lambda_i + \frac{S_i}{\lambda_i} \right\} + \text{const} \tag{18}$$

where

$$S_i = \frac{1}{N} \sum_{n=1}^{N} y_{ni}^2, \qquad y_{ni} = \mathbf{u}_i^{\mathrm{T}} \left(\mathbf{x}_n - \boldsymbol{\mu}\right). \tag{19}$$

Although the log likelihood (18) is not a concave function of $\boldsymbol{\Sigma}$ (one can easily see this by considering the univariate case), one can observe that $\ln p\left(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \to -\infty$ if $\boldsymbol{\Sigma}$ approaches the boundary of the space of symmetric positive-definite matrices, i.e., if $\lambda_i \to 0$ or $\lambda_i \to \infty$ for any i. Therefore, if (17) has a unique solution $\boldsymbol{\Sigma}_{\mathrm{ML}} \succ 0$, then $\boldsymbol{\mu}_{\mathrm{ML}}$ and $\boldsymbol{\Sigma}_{\mathrm{ML}}$ jointly maximize the likelihood.

Note that a similar observation holds when we maximize the log likelihood in terms of the precision $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$, in which case the corresponding log likelihood for $\boldsymbol{\Lambda}$ is given by

$$\ln p\left(\mathbf{X}|\boldsymbol{\mu}_{\mathrm{ML}}, \boldsymbol{\Lambda}\right) = \frac{N}{2} \ln |\boldsymbol{\Lambda}| - \frac{N}{2} \operatorname{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{ML}} \boldsymbol{\Lambda}\right) + \text{const.} \tag{20}$$

Setting the derivative of (20) with respect to $\boldsymbol{\Lambda}$ equal to zero, we indeed obtain $\boldsymbol{\Lambda}_{\mathrm{ML}} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{-1}$. One can also see that (20) is actually a strictly concave function of $\boldsymbol{\Lambda}$ due to the strict concavity of $\ln |\boldsymbol{\Lambda}|$ for $\boldsymbol{\Lambda} \succ 0$ (Magnus and Neudecker, 2007) together with the linearity of $\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{ML}} \boldsymbol{\Lambda}\right)$. See Anderson and Olkin (1985) for further discussions.

Page 100

Equations (2.147) and (2.148): In addition to the mean $\mathbb{E}[\lambda]$ and the variance $\operatorname{var}[\lambda]$, given by (2.147) and (2.148), respectively, we are also interested in the *log expectation* $\mathbb{E}\left[\ln \lambda\right]$, given by (B.30), of the gamma distribution (2.146), which is necessary to evaluate the entropy $\mathrm{H}\left[\lambda\right]$, given by (B.31). Note that the log expectation of the Dirichlet distribution (2.38) is derived in Exercise 2.11 by differentiating its probability with respect to the parameters (the mean and the covariance are concerned in Exercise 2.10). Applying this technique of differentiation, we can calculate the log expectation of the gamma distribution. Here, I would like to state the technique in more general terms (see Section 2.4 for even more general exposition in terms of the *exponential family*), after which we show (B.30). We also find an alternative form of the log expectation $\mathbb{E}\left[\ln \lambda\right]$ in terms of the logarithm of the mean $\ln \mathbb{E}[\lambda]$ and an interesting function related to the digamma function, namely, the *log minus digamma function*.

Score function For a correctly normalized probability distribution $p\left(\mathbf{x}|\boldsymbol{\theta}\right)$ over some random variables $\mathbf{x}$ parameterized by parameters $\boldsymbol{\theta}$ and differentiable with respect to $\boldsymbol{\theta}$,

let us consider how $p(\mathbf{x}|\boldsymbol{\theta})$ changes under perturbations in $\boldsymbol{\theta}$. Specifically, the first-order relative difference in the direction $\boldsymbol{\eta}$ is given by

$$\frac{1}{p(\mathbf{x}|\boldsymbol{\theta})} \lim_{\epsilon \to 0} \left\{ \frac{p(\mathbf{x}|\boldsymbol{\theta} + \epsilon\boldsymbol{\eta}) - p(\mathbf{x}|\boldsymbol{\theta})}{\epsilon} \right\} = \boldsymbol{\eta}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) \tag{21}$$

where we have assumed that $p(\mathbf{x}|\boldsymbol{\theta})$ remains correctly normalized under sufficiently small perturbations in $\boldsymbol{\theta}$; and defined the *score function*, denoted by $\mathbf{g}(\boldsymbol{\theta}, \mathbf{x})$, as the derivative of the log probability with respect to $\boldsymbol{\theta}$ so that

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) \equiv \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}|\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x}|\boldsymbol{\theta})}. \tag{22}$$

Note that the score function (22) is called the Fisher score (6.32) in PRML. In fact, the first-order relative difference (21) is zero on average in whatever the direction $\boldsymbol{\eta}$ because the expectation of the score function (22) vanishes so that

$$\mathbb{E}_{\mathbf{x}} [\mathbf{g}(\boldsymbol{\theta}, \mathbf{x})|\boldsymbol{\theta}] = \mathbf{0} \tag{23}$$

where $\mathbb{E}_{\mathbf{x}} [\cdot|\boldsymbol{\theta}]$ denotes the *conditional expectation* (1.37) so that the above expectation is taken with respect to $p(\mathbf{x}|\boldsymbol{\theta})$.

We can show the general identity (23) by differentiating the both sides of the integral identity

$$\int p(\mathbf{x}|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x} = 1 \tag{24}$$

with respect to $\boldsymbol{\theta}$, giving

$$\nabla_{\boldsymbol{\theta}} \int p(\mathbf{x}|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x} = \mathbf{0} \tag{25}$$

$$\int \nabla_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x} = \mathbf{0} \tag{26}$$

$$\int p(\mathbf{x}|\boldsymbol{\theta}) \, \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x} = \mathbf{0} \tag{27}$$

where we have assumed that we can interchange the order of the derivative and the integral; and used the *log derivative identity*

$$\nabla f = f \, \nabla \ln f. \tag{28}$$

Although we have assumed here that the variables $\mathbf{x}$ are continuous, the same discussion holds if some or all of $\mathbf{x}$ are discrete by replacing the integrations with summations as required.

At this moment, I would like to point out a subtlety in the identity (23). Recall that, when we introduce the score function (22), we have assumed that sufficiently small perturbations in $\boldsymbol{\theta}$ do not affect the correct normalization of $p(\mathbf{x}|\boldsymbol{\theta})$. This assumption is required to show (23) because otherwise the right hand side of (25) would not vanish. Let us take the *multinoulli* distribution $\mathrm{Mult}(\mathbf{x}|\boldsymbol{\mu})$ defined by (142) as an example. Since we cannot change a single parameter μ_k (the normalized probability of observing $x_k = 1$) independently of the others μ_j where $j \neq k$ due to the sum-to-one constraint $\sum_k \mu_k = 1$, it is *not* valid to substitute $\nabla_{\mu_k} \ln \mathrm{Mult}(\mathbf{x}|\boldsymbol{\mu})$ into (23). Instead, we should consider the derivatives with respect to

independent parameters. The unnormalized probabilities $\widetilde{\mu}_k$ related to μ_k through (144) are among such parameters; the corresponding score function is given by

$$\nabla_{\widetilde{\mu}_k} \ln \mathrm{Mult}\left(\mathbf{x}|\boldsymbol{\mu}\right) = \frac{x_k}{\widetilde{\mu}_k} - \frac{1}{\sum_j \widetilde{\mu}_j}. \tag{29}$$

Substituting (29) into (23), we indeed obtain a valid result that $\mathbb{E}[x_k] = \mu_k$.

Log expectation of gamma distribution Now, let us return to the gamma distribution (2.146). The derivative of the log probability with respect to a is given by

$$\frac{\partial}{\partial a} \ln \mathrm{Gam}\left(\lambda|a,b\right) = \ln \lambda - \psi(a) + \ln b \tag{30}$$

where $\psi(\cdot)$ is the *digamma function* given by (66). Substituting (30) into (23), we obtain

$$\mathbb{E}\left[\ln \lambda\right] = \psi(a) - \ln b \tag{31}$$

showing (B.30). Note that one can reproduce the result (2.147) for the mean $\mathbb{E}\left[\lambda\right]$ by substituting the derivative of the log probability with respect to b into (23).

Log minus digamma function It follows from Jensen's inequality (2) that the log expectation $\mathbb{E}\left[\ln \lambda\right]$ is less than the logarithm of the mean $\ln \mathbb{E}\left[\lambda\right]$ because $\ln \xi$ is strictly concave where $\xi > 0$ so that we have

$$\mathbb{E}\left[\ln \lambda\right] < \ln \mathbb{E}\left[\lambda\right] = \ln \frac{a}{b}. \tag{32}$$

The difference between $\ln \mathbb{E}\left[\lambda\right]$ and $\mathbb{E}\left[\ln \lambda\right]$ can be evaluated analytically in this case so that

$$\mathbb{E}\left[\ln \lambda\right] = \ln \mathbb{E}\left[\lambda\right] - \varphi(a) \tag{33}$$

where $\varphi(a) > 0$ is what we call the *log minus digamma function* defined by

$$\varphi(a) \equiv \ln a - \psi(a). \tag{34}$$

The log minus digamma function (34) naturally arises also in deriving the maximum likelihood solution for the gamma distribution as we shall see shortly.

Page 102

Equation (2.155): Although an interpretation for the parameters of the gamma distribution (2.146) has been given, no such an interpretation for the parameters of the Wishart distribution (2.155) is given here nor in Exercise 2.45. Generally speaking, when we construct a probabilistic model with priors, we must choose some reasonable (initial) values for their parameters, known as *hyperparameters*; this calls for an intuitive interpretation for the parameters of such priors. We can give an interpretation for the parameters of the Wishart distribution as follows.

Interpreting parameters of Wishart Let us consider a simple Bayesian inference problem in which, given a set of N observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ for a zero-mean Gaussian random variable, we infer the covariance matrix $\boldsymbol{\Sigma}$ or, equivalently, the precision matrix $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$. The likelihood $p(\mathbf{X}|\boldsymbol{\Lambda})$ in terms of the precision $\boldsymbol{\Lambda}$ is given by

$$p(\mathbf{X}|\boldsymbol{\Lambda}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\Lambda}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n \middle| \mathbf{0}, \boldsymbol{\Lambda}^{-1}\right). \tag{35}$$

If we choose the prior $p(\boldsymbol{\Lambda})$ over $\boldsymbol{\Lambda}$ to be a Wishart distribution so that

$$p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}_0, \nu_0) \tag{36}$$

our analysis can be simplified because it is the conjugate prior. In fact, the posterior $p(\boldsymbol{\Lambda}|\mathbf{X})$ is given by

$$p(\boldsymbol{\Lambda}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\Lambda})\, p(\boldsymbol{\Lambda}) \tag{37}$$

$$\propto |\boldsymbol{\Lambda}|^{N/2} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N} \mathbf{x}_n^{\mathrm{T}} \boldsymbol{\Lambda} \mathbf{x}_n\right\} |\boldsymbol{\Lambda}|^{(\nu_0 - D - 1)/2} \exp\left\{-\frac{1}{2}\operatorname{Tr}\left(\mathbf{W}_0^{-1}\boldsymbol{\Lambda}\right)\right\} \tag{38}$$

$$= |\boldsymbol{\Lambda}|^{(\nu_N - D - 1)/2} \exp\left\{-\frac{1}{2}\operatorname{Tr}\left(\mathbf{W}_N^{-1}\boldsymbol{\Lambda}\right)\right\} \tag{39}$$

where

$$\nu_N = \nu_0 + N \tag{40}$$

$$\mathbf{W}_N^{-1} = \mathbf{W}_0^{-1} + \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}. \tag{41}$$

Reinstating the normalization constant, we indeed see that the posterior becomes again a Wishart distribution of the form

$$p(\boldsymbol{\Lambda}|\mathbf{X}) = \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}_N, \nu_N). \tag{42}$$

This result suggests us how we can interpret the parameters of the Wishart distribution (2.155), namely the scale matrix $\mathbf{W}$ and the number of degrees of freedom ν. Since observing N data points increases the number of degrees of freedom ν by N, we can interpret ν_0 in the prior (36) as the number of "effective" prior observations. The N observations also contribute $N\boldsymbol{\Sigma}_{\mathrm{ML}}$ to the inverse of the scale matrix $\mathbf{W}$ where $\boldsymbol{\Sigma}_{\mathrm{ML}}$ is the maximum likelihood estimate for the covariance of the observations given by

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}. \tag{43}$$

This suggests an interpretation of $\mathbf{W}$ in terms of the "covariance" parameter

$$\boldsymbol{\Sigma} \equiv (\nu\mathbf{W})^{-1}. \tag{44}$$

More specifically, we can interpret $\boldsymbol{\Sigma}_0 = (\nu_0\mathbf{W}_0)^{-1}$ as the covariance of the ν_0 "effective" prior observations. Note that this interpretation is in accordance with another observation that the expectation of $\boldsymbol{\Lambda}$ with respect to the prior (36) is indeed given by $\mathbb{E}[\boldsymbol{\Lambda}] = \nu_0\mathbf{W}_0 = \boldsymbol{\Sigma}_0^{-1}$ where we have used (B.80).

Page 102

Equation (2.157): Again, no interpretation is given for the parameters of the Gaussian-Wishart distribution (2.157) nor for those of the Gaussian-gamma distribution (2.154). Since the Gaussian-gamma can be obtained as a special case of the Gaussian-Wishart where the dimension is one so that $D = 1$, we shall make an interpretation only for the parameters of the Gaussian-Wishart here.

Interpreting parameters of Gaussian-Wishart Let us consider a problem of inferring the mean $\boldsymbol{\mu}$ and the precision $\boldsymbol{\Lambda}$ given the Gaussian likelihood

$$p\left(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Lambda}\right) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \tag{45}$$

and the Gaussian-Wishart prior

$$p\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}\right) = \mathcal{N}\left(\boldsymbol{\mu} \middle| \boldsymbol{\mu}_0, \left(\beta_0 \boldsymbol{\Lambda}\right)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}|\mathbf{W}_0, \nu_0\right). \tag{46}$$

At this moment, we introduce notations for the maximum likelihood estimates for the mean and the covariance given the N observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, i.e.,

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n, \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \tag{47}$$

respectively. Evaluating the posterior, we have

$$\begin{aligned}
p\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}\right) &\propto p\left(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Lambda}\right) p\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}\right) && (48)\\
&\propto |\boldsymbol{\Lambda}|^{N/2} \exp\left\{-\frac{1}{2} \sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Lambda} \left(\mathbf{x}_n - \boldsymbol{\mu}\right)\right\} \\
&\quad \times |\boldsymbol{\Lambda}|^{(\nu_0 - D)/2} \exp\left\{-\frac{1}{2} \mathrm{Tr}\left(\left\{\mathbf{W}_0^{-1} + \beta_0 \left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^{\mathrm{T}}\right\} \boldsymbol{\Lambda}\right)\right\} && (49)\\
&= |\boldsymbol{\Lambda}|^{(\nu_N - D)/2} \exp\left\{-\frac{1}{2} \mathrm{Tr}\left(\left\{\mathbf{W}_N^{-1} + \beta_N \left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)\left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)^{\mathrm{T}}\right\} \boldsymbol{\Lambda}\right)\right\} && (50)
\end{aligned}$$

where[6]

$$\begin{aligned}
\beta_N &= \beta_0 + N && (55)\\
\beta_N \boldsymbol{\mu}_N &= \beta_0 \boldsymbol{\mu}_0 + N \boldsymbol{\mu}_{\mathrm{ML}} && (56)\\
\nu_N &= \nu_0 + N && (57)\\
\mathbf{W}_N^{-1} &= \mathbf{W}_0^{-1} + N \left[\boldsymbol{\Sigma}_{\mathrm{ML}} + \frac{\beta_0}{\beta_N} \left(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}_0\right)^{\mathrm{T}}\right]. && (58)
\end{aligned}$$

[6]The form (58) of $\mathbf{W}_N^{-1}$ is a little tricky to obtain so that I would like to show a more detailed derivation here. Collecting and evaluating the coefficients of $\boldsymbol{\Lambda}$ inside $\mathrm{Tr}(\cdot)$ in the posterior (49), we have

$$\begin{aligned}
&\mathbf{W}_0^{-1} + \sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^{\mathrm{T}} + \beta_0 \left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^{\mathrm{T}} && (51)\\
&= \mathbf{W}_0^{-1} + N \boldsymbol{\Sigma}_{\mathrm{ML}} + N \left(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathrm{ML}}\right)\left(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathrm{ML}}\right)^{\mathrm{T}} + \beta_0 \left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^{\mathrm{T}} && (52)\\
&= \mathbf{W}_0^{-1} + N \boldsymbol{\Sigma}_{\mathrm{ML}} + \beta_N \left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)\left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)^{\mathrm{T}} - \beta_N \boldsymbol{\mu}_N \boldsymbol{\mu}_N^{\mathrm{T}} + N \boldsymbol{\mu}_{\mathrm{ML}} \boldsymbol{\mu}_{\mathrm{ML}}^{\mathrm{T}} + \beta_0 \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^{\mathrm{T}} && (53)\\
&= \mathbf{W}_0^{-1} + N \boldsymbol{\Sigma}_{\mathrm{ML}} + \beta_N \left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)\left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)^{\mathrm{T}} + \frac{\beta_0 N}{\beta_0 + N} \left(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}_0\right)^{\mathrm{T}}. && (54)
\end{aligned}$$

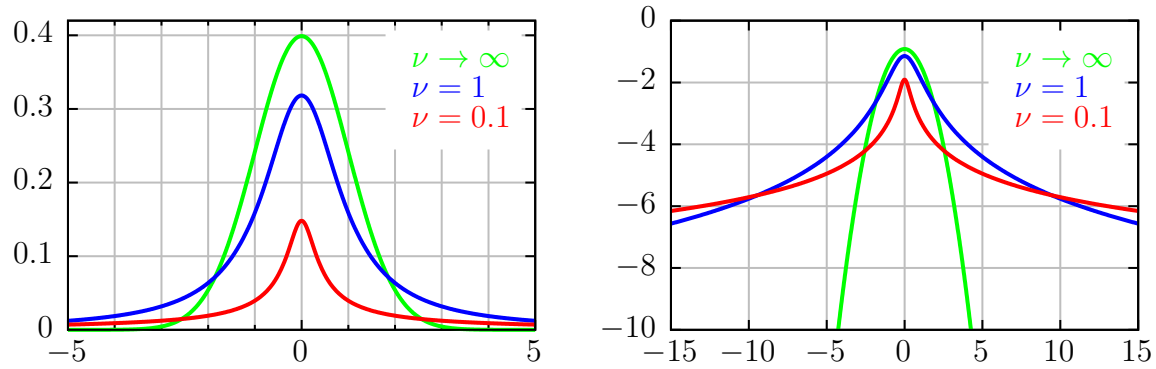

Figure 3 Plot of Student's t-distribution density functions $\mathrm{St}\left(x|\mu,\lambda,\nu\right)$ (left) and corresponding log density functions $\ln \mathrm{St}\left(x|\mu,\lambda,\nu\right)$ (right) for various values of ν where we have fixed $\mu = 0$ and $\lambda = 1$.

Thus, we find that the posterior is again a Gaussian-Wishart of the form

$$p\left(\boldsymbol{\mu},\boldsymbol{\Lambda}|\mathbf{X}\right) = \mathcal{N}\left(\boldsymbol{\mu}\middle|\boldsymbol{\mu}_N,\left(\beta_N\boldsymbol{\Lambda}\right)^{-1}\right)\mathcal{W}\left(\boldsymbol{\Lambda}|\mathbf{W}_N,\nu_N\right). \tag{59}$$

Note that a similar result is obtained in Section 10.2.1 for a Bayesian mixture of Gaussians model in which we assume a Gaussian-Wishart prior for each Gaussian component.

From the above result, we see that the parameters β_0 and $\boldsymbol{\mu}_0$ for the mean $\boldsymbol{\mu}$ can be interpreted somewhat independently of those ν_0 and $\mathbf{W}_0$ for the precision $\boldsymbol{\Lambda}$. We can interpret β_0 as the number of "effective" prior observations for $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_0$ as the mean of the β_0 prior observations. The interpretation of ν_0 and $\mathbf{W}_0$ is similar to the one we have made in the previous erratum except that we have in (58) a term due to the uncertainty in $\boldsymbol{\mu}$, that is, a term involving the outer product of the difference between the maximum likelihood mean $\boldsymbol{\mu}_{\mathrm{ML}}$ and the prior mean $\boldsymbol{\mu}_0$, scaled by β_0/β_N.

Page 102

Paragraph –1, Line –2: "Gamma" should read "gamma" (without capitalization).

Page 103

Figure 2.15: The tails of Student's t-distributions are too high; one can easily see that, if compared to the corresponding Gaussian distribution labeled $\nu \to \infty$, the t-distributions are not correctly normalized. Figure 3 gives the correct plot.

Page 103

Paragraph –1, Line –3: As pointed out in the text, the maximum likelihood solution for Student's t-distribution can be most easily found by the *expectation maximization* (EM) algorithm, which we study for discrete and continuous latent variables in Chapters 9 and 12, respectively; it is not until Exercise 12.24 that we apply the EM algorithm to the problem of maximum likelihood for the (multivariate) Student's t-distribution (2.162). Although we have to defer the derivation of the above mentioned EM algorithm for some time, it is useful to have considered a related problem of maximum likelihood for the gamma distribution (2.146) here in advance, because, since the t-distribution is obtained by marginalizing over a gamma

distributed precision as we have seen in (2.158), we need to estimate the gamma distribution as a subproblem of the EM for the t-distribution.

Maximum likelihood for gamma distribution Given a data set $\mathbf{x} = \{x_1, \ldots, x_N\}$, we consider a likelihood function of the form

$$p(\mathbf{x}|a, b) = \prod_{n=1}^{N} \mathrm{Gam}(x_n|a, b) \tag{60}$$

where the gamma distribution $\mathrm{Gam}(\lambda|a, b)$ is given by (2.146). The log likelihood is given by

$$\ln p(\mathbf{x}|a, b) = N\{-\ln \Gamma(a) + a \ln b + (a-1) \ln \widehat{x} - b\overline{x}\} \tag{61}$$

where $\overline{x}$ and $\widehat{x}$ denote the *arithmetic mean* and the *geometric mean*, respectively, so that

$$\overline{x} = \frac{1}{N} \sum_{n=1}^{N} x_n, \qquad \ln \widehat{x} = \frac{1}{N} \sum_{n=1}^{N} \ln x_n. \tag{62}$$

We see from (61) that $\overline{x}$ and $\ln \widehat{x}$ are the sufficient statistics of the gamma distribution. Here, we assume that $x_n > 0$ for all n (which holds with probability one if x_n has been drawn from a gamma distribution) so that we have $\overline{x} > 0$ and $\widehat{x} > 0$.

Let us first assume that $a > 0$ is known. It is easy to see that the log likelihood (61) is a strictly concave function of $b > 0$. We can maximize the likelihood by setting the derivative of (61) with respect to b equal to zero, which gives $b = a/\overline{x}$. Back substituting this into (61), we have

$$\ln p(\mathbf{x}|a, b)|_{b=a/\overline{x}} = N\{-\ln \Gamma(a) + a \ln a - a \ln \overline{x} + (a-1) \ln \widehat{x} - a\}. \tag{63}$$

Next we maximize (63) with respect to $a > 0$. This can be done by setting the derivative of (63) with respect to a equal to zero, which gives a nonlinear equation of the form

$$\varphi(a) = \ln \overline{x} - \ln \widehat{x} \tag{64}$$

where $\varphi(\cdot)$ is the *log minus digamma function* given by (34). One can see that (63) is again a strictly concave function of $a > 0$ because $\varphi(a)$ is a strictly monotonically decreasing function so that $\varphi'(a) < 0$.

It follows from Jensen's inequality (2) that $\ln \overline{x} \geqslant \ln \widehat{x}$ (which implies $\overline{x} \geqslant \widehat{x}$ because of the monotonicity of the logarithm). Here, we further assume that the strict inequality $\ln \overline{x} > \ln \widehat{x}$ holds so that the right hand side of (64) lies among $(0, \infty)$. Since the log minus digamma function $\varphi : (0, \infty) \to (0, \infty)$ is bijective and thus has the inverse function $\varphi^{-1} : (0, \infty) \to (0, \infty)$, we can solve (64) uniquely for $a > 0$. Substituting this into $b = a/\overline{x}$, we finally obtain the maximum likelihood solution for the gamma distribution

$$a_{\mathrm{ML}} = \varphi^{-1}(\ln \overline{x} - \ln \widehat{x}), \qquad b_{\mathrm{ML}} = \frac{a_{\mathrm{ML}}}{\overline{x}}. \tag{65}$$

Page 104

Paragraph 1, Line 4: In practical applications, the importance of *robustness to outliers* cannot be overemphasized. Here, I would like to point out that, particularly in the context of robust

regression, there exist historically a number of heuristic approaches to robustness such as *M-estimators* (Press et al., 1992; Szeliski, 2010), in which the standard least squares method is modified so as to use a more "robust" cost function. In this respect, the robust regression in terms of Student's t-distribution can be regarded as an M-estimator where the cost function is derived from its negative log likelihood.

An M-estimator can be solved iteratively by approximating the cost functions successively in terms of quadratic bounds. Called *iteratively reweighted least squares* or IRLS, this algorithm closely resembles the EM algorithm. In fact, one can identify the IRLS and the EM for the robust regression in terms of Student's t-distribution. Note also that the successive quadratic approximation in IRLS can be regarded as a *local variational method* discussed in Chapter 10.

Although we are free to choose from a broad class of cost functions in M-estimators, such a heuristic choice makes our Bayesian analysis difficult. For instance, M-estimators need a separate evaluation data set for selecting hyperparameters. On the other hand, probabilistic models such as the robust regression model in terms of Student's t-distribution allow us to perform model selection in a consistent way without the need for an evaluation data set, while they, in principle, do not suffer from overfitting.

Page 104

The text after (2.160): The Gaussian $\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}\right)$ should read $\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right)$.

Page 116

Equation (2.224): The right hand side should be a zero vector $\mathbf{0}$ instead of a scalar zero 0.

Page 126

The caption of Figure 2.28: The red, green, and blue points correspond to the "homogeneous," "annular," and "laminar" (or "stratified") classes, respectively.

Page 129

Exercise 2.9, Line 1: Remove the period (.) after **www**.

Page 130

Equation (2.277): In order to be consistent with the mathematical notation in PRML, the differential operator d should be an upright d. Specifically, the *digamma function* is given by

$$\psi(a) \equiv \frac{\mathrm{d}}{\mathrm{d}a} \ln \Gamma(a) = \frac{\Gamma'(a)}{\Gamma(a)}. \tag{66}$$

Note that the digamma function (66) is also known as the *psi function* (Abramowitz and Stegun, 1964; Olver et al., 2016).

Page 138

Equation (3.1): The lower ellipsis (. . .) should be centered ($\cdots$).

Page 141

Equation (3.13): The use of the gradient operator ∇ is not consistent here. As in (2.224), the gradient of a scalar function is usually defined as a column vector of derivatives so that (3.13) should read[7]

$$\nabla \ln p\left(\mathbf{t}|\mathbf{w},\beta\right) = \beta \sum_{n=1}^{N} \left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\left(\mathbf{x}_n\right)\right\} \boldsymbol{\phi}\left(\mathbf{x}_n\right). \tag{67}$$

Moreover, I would like to also suggest that we should give a definition for the gradient before we use it or, at least, in an appendix. Although Appendix C defines the vector derivative $\frac{\partial}{\partial \mathbf{x}}$, which is used interchangeably with the gradient $\nabla_{\mathbf{x}}$ throughout PRML, there is no mention of the gradient. We shall come back to this issue later in this report.

Page 142

Equation (3.14): The left hand side should be a zero vector $\mathbf{0}$ instead of a scalar zero 0 so that (3.14) should read

$$\mathbf{0} = \sum_{n=1}^{N} t_n \boldsymbol{\phi}\left(\mathbf{x}_n\right) - \left(\sum_{n=1}^{N} \boldsymbol{\phi}\left(\mathbf{x}_n\right) \boldsymbol{\phi}\left(\mathbf{x}_n\right)^{\mathrm{T}}\right) \mathbf{w} \tag{68}$$

where we have used the gradient of the form (67) instead of (3.13).

Page 146

Equation (3.31): The left hand side should be $\mathbf{y}\left(\mathbf{x},\mathbf{W}\right)$ instead of $\mathbf{y}\left(\mathbf{x},\mathbf{w}\right)$.

Page 166

Paragraph 2, Line 1: "Gamma" should read "gamma" (without capitalization).

Pages 168–169, and 177

Equations (3.88), (3.93), and (3.117) as well as the text before (3.93): The derivative operators should be partial differentials. For example, (3.117) should read

$$\frac{\partial}{\partial \alpha} \ln \left|\mathbf{A}\right| = \mathrm{Tr}\left(\mathbf{A}^{-1} \frac{\partial}{\partial \alpha} \mathbf{A}\right). \tag{69}$$

Page 170

Figure 3.15: The eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ are unit vectors so that their orientations should be shown as in Figure 2.7 on Page 81. Or, the scaled vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ should be labeled as $\lambda_1^{-1/2}\mathbf{u}_1$ and $\lambda_2^{-1/2}\mathbf{u}_2$, respectively.

Page 179

Paragraph 1, Line –4: The decision surfaces are defined by linear *equations* of the input vector $\mathbf{x}$ and thus are $(D-1)$-dimensional hyperplanes within the D-dimensional input space.

[7] Note that we use a different typeface (from a D-dimensional target variable $\mathbf{t}$) for the N-dimensional data vector $\mathbf{t} = \left(t_1,\ldots,t_N\right)^{\mathrm{T}}$ consisting of one-dimensional target variables $\{t_n\}$. See "Mathematical Notation" for PRML on Pages xi–xii.

Page 190

Equation (4.33): The right hand side should be a zero vector $\mathbf{0}$ instead of a scalar zero 0.

Page 205

Equation (4.88): The differential operator d should be an upright d.

Page 207

Equation (4.92): The gradient and the Hessian in the right hand side, which are in general functions of the parameter $\mathbf{w}$, must be evaluated at the previous estimate $\mathbf{w}^{\text{old}}$ for the parameter. Thus, (4.92) should read

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \left[\mathbf{H}\left(\mathbf{w}^{\text{old}}\right)\right]^{-1} \nabla E\left(\mathbf{w}^{\text{old}}\right) \tag{70}$$

where $\mathbf{H}(\mathbf{w}) \equiv \nabla\nabla E(\mathbf{w})$ is the Hessian matrix whose elements comprise the second derivatives of $E(\mathbf{w})$ with respect to the components of $\mathbf{w}$.

Page 210

Equation (4.110) and the preceding text: The left hand side of (4.110) is obtained by taking the gradient of $\nabla_{\mathbf{w}_j} E$ given in (4.109) with respect to $\mathbf{w}_k$ and corresponds to the (k, j)-th block of the Hessian, *not* the (j, k)-th. Thus, (4.110) should read

$$\nabla_{\mathbf{w}_k}\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N} y_{nj}\left(I_{kj} - y_{nk}\right)\boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}. \tag{71}$$

To be clear, we have used the following notation. If we group all the parameters $\mathbf{w}_1, \ldots, \mathbf{w}_K$ into a column vector

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_K \end{pmatrix} \tag{72}$$

the gradient and the Hessian of the error function $E(\mathbf{w})$ with respect to $\mathbf{w}$ are given by

$$\nabla_{\mathbf{w}} E = \begin{pmatrix} \nabla_{\mathbf{w}_1} E \\ \vdots \\ \nabla_{\mathbf{w}_K} E \end{pmatrix}, \qquad \nabla_{\mathbf{w}}\nabla_{\mathbf{w}} E = \begin{pmatrix} \nabla_{\mathbf{w}_1}\nabla_{\mathbf{w}_1} E & \cdots & \nabla_{\mathbf{w}_1}\nabla_{\mathbf{w}_K} E \\ \vdots & \ddots & \vdots \\ \nabla_{\mathbf{w}_K}\nabla_{\mathbf{w}_1} E & \cdots & \nabla_{\mathbf{w}_K}\nabla_{\mathbf{w}_K} E \end{pmatrix} \tag{73}$$

respectively.

Pages 212–214

Equations (4.119), (4.122), (4.126), and (4.128): The differential operator d should be an upright d.

Page 237

Equation (5.26): The right hand side should be a zero vector $\mathbf{0}$ instead of a scalar zero 0.

Page 237

Paragraph 4, Line 2: $\nabla E(\mathbf{w}) = 0$ should read $\nabla E(\mathbf{w}) = \mathbf{0}$ (the right hand side should be a zero vector $\mathbf{0}$).

Page 238

Paragraph 2, Line 3: $\nabla E(\mathbf{w}) = 0$ should read $\nabla E(\mathbf{w}) = \mathbf{0}$ (the right hand side should be a zero vector $\mathbf{0}$).

Page 239

Figure 5.6: The eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ are unit vectors so that their orientations should be shown as in Figure 2.7 on Page 81. Or, the scaled vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ should be labeled as $\lambda_1^{-1/2}\mathbf{u}_1$ and $\lambda_2^{-1/2}\mathbf{u}_2$, respectively.

Page 251

Paragraph 2, Line 1: The outer product approximation to the Hessian of the form (5.84) is usually referred to as the *Gauss-Newton* approximation (Press et al., 1992), which not only eliminates the computation of second derivatives but also guarantees that the Hessian thus approximated is positive (semi)definite, whereas the *Levenberg-Marquardt* method (Press et al., 1992) is a method that improves the numerical stability of (Gauss-)Newton type iterations by correcting the Hessian matrix so as to be more diagonal dominant. Let us now compare the two types of approximation to the Hessian, i.e., Gauss-Newton and Levenberg-Marquardt, more specifically in the following. We first observe that the Gauss-Newton approximation to the Hessian given in the right hand side of (5.84) can be written succinctly in terms of matrix product as

$$\mathbf{H}_{\text{GN}} = \mathbf{J}^{\text{T}}\mathbf{J} \tag{74}$$

where $\mathbf{J} = (\nabla a_1, \ldots, \nabla a_N)^{\text{T}}$ is the Jacobian of the activations $a_1, \ldots, a_N$ with respect to the parameters (weights and biases). The Levenberg-Marquardt approximation to the above Hessian typically takes the form

$$\mathbf{H}_{\text{LM}} = \mathbf{J}^{\text{T}}\mathbf{J} + \lambda\mathbf{I} \tag{75}$$

or

$$\mathbf{H}_{\text{LM}} = \mathbf{J}^{\text{T}}\mathbf{J} + \lambda\,\text{diag}\left(\mathbf{J}^{\text{T}}\mathbf{J}\right) \tag{76}$$

where we have introduced an adjustable damping factor $\lambda \geqslant 0$ (which will be adjusted through the iterations) and defined that, for a square matrix $\mathbf{A} = (A_{ij})$, $\text{diag}(\mathbf{A})$ is a diagonal matrix obtained by setting the off-diagonal elements equal to zero so that $\text{diag}(\mathbf{A}) = \text{diag}(A_{ii})$.

Page 259

Paragraph 1, Line –1: The parameters rescaling should be $\lambda_1 \to a^2\lambda_1$ and $\lambda_2 \to c^{-2}\lambda_2$.

Page 266

Equation (5.132): The third occurrence of the superscript T for matrix transpose should be an upright T.

Page 267

Equation (5.134): The superscript T should be an upright T.

Page 275

The text after (5.154): The identity matrix $\mathbf{I}$ should multiply $\sigma_k^2(\mathbf{x}_n)$.

Page 277

Equation (5.160): The factor L should multiply $\sigma_k^2(\mathbf{x})$ because we have

$$s^2(\mathbf{x}) = \mathbb{E}\left[\mathrm{Tr}\left\{(\mathbf{t} - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right])\left(\mathbf{t} - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\right)^{\mathrm{T}}\right\}\middle|\mathbf{x}\right] \tag{77}$$

$$= \sum_{k=1}^{K} \pi_k(\mathbf{x})\,\mathrm{Tr}\left\{\sigma_k^2(\mathbf{x})\mathbf{I} + \left(\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\right)\left(\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\right)^{\mathrm{T}}\right\} \tag{78}$$

$$= \sum_{k=1}^{K} \pi_k(\mathbf{x})\left\{L\sigma_k^2(\mathbf{x}) + \left\|\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\right\|^2\right\} \tag{79}$$

where L is the dimensionality of $\mathbf{t}$.

Page 279

Equations (5.169) and (5.171): The superscripts $\mathbf{T}$ (in a bold typeface) should read T (in a roman typeface).

Page 295

Paragraph 1, Line 1: The vector $\mathbf{x}$ should be a column vector so that $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$.

Page 318

Equations (6.93) and (6.94) as well as the text before (6.93): The text and the equations should read: We can evaluate the derivative of $a_n^\star$ with respect to θ_j by differentiating the relation (6.84) with respect to θ_j to give

$$\frac{\partial \mathbf{a}_N^\star}{\partial \theta_j} = \frac{\partial \mathbf{C}_N}{\partial \theta_j}\left(\mathbf{t}_N - \boldsymbol{\sigma}_N\right) - \mathbf{C}_N\mathbf{W}_N\frac{\partial \mathbf{a}_N^\star}{\partial \theta_j} \tag{80}$$

where the derivatives are Jacobians defined by (C.16) for a vector and analogously by (189) for a matrix. Rearranging (80) then gives

$$\frac{\partial \mathbf{a}_N^\star}{\partial \theta_j} = \left(\mathbf{I} + \mathbf{C}_N\mathbf{W}_N\right)^{-1}\frac{\partial \mathbf{C}_N}{\partial \theta_j}\left(\mathbf{t}_N - \boldsymbol{\sigma}_N\right). \tag{81}$$

Page 333

Equation (7.29): $\frac{\partial L}{\partial \mathbf{w}} = 0$ should read $\nabla_{\mathbf{w}} L = \mathbf{0}$.

Page 335

Paragraph 1, Line 12: The term "protected conjugate gradients" should read "*projected* conjugate gradients."

Page 341

Equation (7.57): $\frac{\partial L}{\partial \mathbf{w}} = 0$ should read $\nabla_{\mathbf{w}} L = \mathbf{0}$.

Page 354

Equation (7.112): The mean $\mathbf{w}^\star$ of the Laplace approximation to the posterior $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha})$ can only be obtained iteratively by, say, IRLS as described in the text so that (7.112) does not represent an explicit solution and is thus best removed.

As we shall see shortly, it is however useful to note that, at the convergence of IRLS, we have the following implicit equations for $\mathbf{w}^\star$

$$\nabla_{\mathbf{w}} \ln p(\mathbf{w}^\star|\mathbf{t}, \boldsymbol{\alpha}) = \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{t} - \mathbf{y}^\star) - \mathbf{A}\mathbf{w}^\star = \mathbf{0} \tag{82}$$

where $\mathbf{y}^\star$ is $\mathbf{y}$ evaluated at $\mathbf{w} = \mathbf{w}^\star$ so that

$$\mathbf{y}^\star = \mathbf{y}|_{\mathbf{w}=\mathbf{w}^\star}\,. \tag{83}$$

Page 354

Equation (7.113): Let us note that the precision (the inverse of the covariance $\boldsymbol{\Sigma}$) of the Laplace approximation to the posterior $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha})$ is given by the Hessian of the negative log posterior evaluated at $\mathbf{w}^\star$ so that

$$\boldsymbol{\Sigma}^{-1} = -\nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \ln p(\mathbf{w}^\star|\mathbf{t}, \boldsymbol{\alpha})\,. \tag{84}$$

The covariance $\boldsymbol{\Sigma}$ should thus be given by

$$\boldsymbol{\Sigma} = \left(\mathbf{A} + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}^\star\boldsymbol{\Phi}\right)^{-1} \tag{85}$$

where $\mathbf{B}^\star$ is $\mathbf{B}$ evaluated at $\mathbf{w}^\star$ so that

$$\mathbf{B}^\star = \mathbf{B}|_{\mathbf{w}=\mathbf{w}^\star}\,. \tag{86}$$

Page 355

Equation (7.117): The typeface of the vector $\mathbf{y}$ in (7.117) should be that in (7.110), i.e., $\mathbf{y}$. Moreover, $\mathbf{B}$ and $\mathbf{y}$ should be those evaluated at $\mathbf{w} = \mathbf{w}^\star$ so that $\widehat{\mathbf{t}}$ is given by

$$\widehat{\mathbf{t}} = \boldsymbol{\Phi}\mathbf{w}^\star + (\mathbf{B}^\star)^{-1}(\mathbf{t} - \mathbf{y}^\star)\,. \tag{87}$$

It should also be noted here that we define $\widehat{\mathbf{t}}$ as (87) in order that we can write the posterior mean $\mathbf{w}^\star$ in terms of $\widehat{\mathbf{t}}$ so that

$$\mathbf{w}^\star = \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}^\star\widehat{\mathbf{t}} \tag{88}$$

because we have

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}^\star\widehat{\mathbf{t}} &= \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}^\star\boldsymbol{\Phi}\mathbf{w}^\star + \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{t} - \mathbf{y}^\star) && (89)\\
&= \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}^\star\boldsymbol{\Phi}\mathbf{w}^\star + \boldsymbol{\Sigma}\mathbf{A}\mathbf{w}^\star && (90)\\
&= \boldsymbol{\Sigma}\left(\mathbf{A} + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}^\star\boldsymbol{\Phi}\right)\mathbf{w}^\star && (91)\\
&= \mathbf{w}^\star && (92)
\end{aligned}$$

where we have made use of (87), (82), and (85). We shall make use of (88) when we analyze the RVM classification problem (see below).

Page 355

Equation (7.118): Although this marginal distribution cannot be obtained directly from the Laplace approximation (7.114) to the marginal $p(\mathbf{t}|\boldsymbol{\alpha})$ of the RVM classification problem, it can be shown to be an (approximate) marginal for a "linearized" version of the classification problem where $\widehat{\mathbf{t}}$, given by (87), serves as the target (Tipping and Faul, 2003). Since the linearized problem can be regarded as an RVM regression problem having data-dependent precisions, we first review such a regression problem, after which we derive the marginal for the linearized classification problem.

RVM regression The likelihood for the RVM regression problem with data-dependent precisions $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_N\}$ is given by

$$p(\mathbf{t}|\mathbf{w}, \boldsymbol{\beta}) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \middle| \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n, \beta_n^{-1}\right) \tag{93}$$

$$= \mathcal{N}\left(\mathbf{t} \middle| \boldsymbol{\Phi}\mathbf{w}, \mathbf{B}^{-1}\right) \tag{94}$$

where we have omitted the conditioning on $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ to keep the notation uncluttered; and written $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ and

$$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}} \tag{95}$$

$$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N)^{\mathrm{T}} \tag{96}$$

$$\mathbf{B} = \mathrm{diag}(\beta_1, \ldots, \beta_N). \tag{97}$$

The prior is the same as (7.80) so that

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}\left(w_i \middle| 0, \alpha_i^{-1}\right) \tag{98}$$

$$= \mathcal{N}\left(\mathbf{w} \middle| \mathbf{0}, \mathbf{A}^{-1}\right) \tag{99}$$

where

$$\mathbf{w} = (w_1, \ldots, w_M)^{\mathrm{T}} \tag{100}$$

$$\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_M). \tag{101}$$

The joint distribution is given by a linear-Gaussian model of the form

$$p(\mathbf{t}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\mathbf{t}|\mathbf{w}, \boldsymbol{\beta})\, p(\mathbf{w}|\boldsymbol{\alpha}) \tag{102}$$

$$= \mathcal{N}\left(\mathbf{t} \middle| \boldsymbol{\Phi}\mathbf{w}, \mathbf{B}^{-1}\right) \mathcal{N}\left(\mathbf{w} \middle| \mathbf{0}, \mathbf{A}^{-1}\right). \tag{103}$$

Making use of the general results (2.115) and (2.116) for the marginal and the conditional Gaussians, we can readily evaluate the marginal and the posterior distributions again as Gaussians. The posterior is given by

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathcal{N}(\mathbf{w}|\mathbf{w}^{\star}, \boldsymbol{\Sigma}) \tag{104}$$

where

$$\mathbf{w}^{\star} = \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}\mathbf{t} \tag{105}$$

$$\boldsymbol{\Sigma} = \left(\mathbf{A} + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\Phi}\right)^{-1}. \tag{106}$$

The marginal is given by

$$p\left(\mathbf{t}|\boldsymbol{\alpha},\beta\right)=\mathcal{N}\left(\mathbf{t}|\mathbf{0},\mathbf{C}\right) \tag{107}$$

where

$$\mathbf{C}=\mathbf{B}^{-1}+\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}. \tag{108}$$

RVM classification Let us now return to the RVM classification problem. We have already seen that the posterior can be approximated by the Laplace approximation so that

$$p\left(\mathbf{w}|\mathbf{t},\boldsymbol{\alpha}\right)\approx\mathcal{N}\left(\mathbf{w}|\mathbf{w}^{\star},\boldsymbol{\Sigma}\right) \tag{109}$$

where the mean $\mathbf{w}^{\star}$ can be obtained by the IRLS algorithm as we have discussed; and the covariance $\boldsymbol{\Sigma}$ is given by (85).

Here, we note that $\mathbf{w}^{\star}$ can be written in the form (88). Comparing (88) with (105) and (85) with (106), we see that the Laplace approximation locally maps the classification problem to a regression problem with the data-dependent precision matrix $\mathbf{B}^{\star}$, given by (86), where the target vector $\mathbf{t}$ is replaced by the "linearized" target $\widehat{\mathbf{t}}$, given by (87).

Assuming that the distribution over $\widehat{\mathbf{t}}$ can be approximated by the Laplace approximation (as we have done in (109) for $\mathbf{w}$); and making use of the linear-Gaussian relation, we can obtain the corresponding marginal for the linearized problem in the form

$$p\left(\widehat{\mathbf{t}}\middle|\boldsymbol{\alpha}\right)\approx\mathcal{N}\left(\widehat{\mathbf{t}}\middle|\mathbf{0},\mathbf{C}\right) \tag{110}$$

where

$$\mathbf{C}=\left(\mathbf{B}^{\star}\right)^{-1}+\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}. \tag{111}$$

The right hand side of (110) takes the same form as the marginal (107) of the regression problem so that "we can apply the same analysis of sparsity and obtain the same fast learning algorithm" (Page 355, Paragraph –4, Line –3).

Page 355

Equation (7.119): Both $\mathbf{A}$ and $\mathbf{B}$ should be inverted and, moreover, $\mathbf{B}$ should be that evaluated at $\mathbf{w}=\mathbf{w}^{\star}$ so that (7.119) should read (111).

Page 414

The caption of Figure 8.53, Line 6: The term "max-product" should be "max-sum."

Page 425

Equation (9.3): The right hand side should be a zero vector $\mathbf{0}$ instead of a scalar zero 0.

Page 432

The text after (9.13): I would like to point out for clarity that the prior $p(\mathbf{z})$ given by (9.10) is a multinomial distribution or, more precisely, a *multinoulli* distribution (142) so that

$$p\left(\mathbf{z}\right)=\mathrm{Mult}\left(\mathbf{z}|\boldsymbol{\pi}\right)=\prod_{k=1}^{K}\pi_{k}^{z_{k}}. \tag{112}$$

Moreover, we see that the posterior $p(\mathbf{z}|\mathbf{x})$ again becomes a multinoulli distribution of the form

$$p(\mathbf{z}|\mathbf{x}) = \text{Mult}(\mathbf{z}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} \gamma_k^{z_k} \tag{113}$$

where we have written $\gamma_k \equiv \gamma(z_k)$. Called the *responsibility*, γ_k is given by (9.13), which can also be found directly by inspecting the functional form of the joint distribution

$$p(\mathbf{z})\, p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}^{z_k} \tag{114}$$

and noting that the multinoulli distribution (142) can be expressed in terms of unnormalized probabilities as shown in (143) where the normalized probabilities are given by (144). This observation helps the reader understand that evaluating the responsibilities γ_k indeed corresponds to the E step of the EM algorithm.

Page 434

Equation (9.15): Although the official errata (Svensén and Bishop, 2011) states that σ_j in the right hand side should be raised to a power of D, the whole right hand side should be raised to D so that (9.15) should read

$$\mathcal{N}\left(\mathbf{x}_n \middle| \mathbf{x}_n, \sigma_j^2 \mathbf{I}\right) = \frac{1}{\left(2\pi\sigma_j^2\right)^{D/2}}. \tag{115}$$

Page 435

Equation (9.16): The left hand side should be a zero vector $\mathbf{0}$ instead of a scalar zero 0.

Page 453

Paragraph 1: The old and new parameters should read $\boldsymbol{\theta}^{\text{old}}$ and $\boldsymbol{\theta}^{\text{new}}$ (without parentheses), respectively, as in (9.74) and the text.

Page 465

Equations (10.6) and (10.7): The lower bound of the form (10.6) for variational Bayes will be later recognized as "a negative Kullback-Leibler divergence between $q_j(\mathbf{Z}_j)$ and $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$" (Page 465, Paragraph –1, Line –2). However, there is no point in taking the Kullback-Leibler divergence between two probability distributions over different sets of random variables; such a quantity is undefined. Moreover, the discussion here seems to be somewhat redundant. Without introducing an intermediate quantity like $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$, we can rewrite (10.6) and (10.7) directly in terms of $q_j^\star(\mathbf{Z}_j)$. Specifically, writing down the terms dependent on one of the factors $q_j(\mathbf{Z}_j)$, we obtain the lower bound $\mathcal{L}(q)$ in the form

$$\mathcal{L}(q) = \int q_j(\mathbf{Z}_j)\, \mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_j}[\ln p(\mathbf{X}, \mathbf{Z})]\, \mathrm{d}\mathbf{Z}_j - \int q_j(\mathbf{Z}_j) \ln q_j(\mathbf{Z}_j)\, \mathrm{d}\mathbf{Z}_j + \text{const} \tag{116}$$

$$= -\,\mathrm{KL}\left(q_j \middle\| q_j^\star\right) + \text{const} \tag{117}$$

where we have assumed that the expectation $\mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_j}[\cdot]$ is taken with respect to $\mathbf{Z}$ but $\mathbf{Z}_j$ so that

$$\mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_j}\left[\ln p\left(\mathbf{X},\mathbf{Z}\right)\right] = \int\cdots\int \ln p\left(\mathbf{X},\mathbf{Z}\right)\prod_{i\neq j} q_i\left(\mathbf{Z}_i\right)\mathrm{d}\mathbf{Z}_i \tag{118}$$

and defined a new distribution $q_j^\star\left(\mathbf{Z}_j\right)$ over $\mathbf{Z}_j$ by the relation

$$\ln q_j^\star\left(\mathbf{Z}_j\right) = \mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_j}\left[\ln p\left(\mathbf{X},\mathbf{Z}\right)\right] + \text{const.} \tag{119}$$

It directly follows from (117) that, since the lower bound $\mathcal{L}(q)$ is the negative Kullback-Leibler divergence between $q_j(\mathbf{Z}_j)$ and $q_j^\star(\mathbf{Z}_j)$ up to some additive constant, the maximum of $\mathcal{L}(q)$ occurs when $q_j(\mathbf{Z}_j) = q_j^\star(\mathbf{Z}_j)$.

Page 465

The text before (10.8): The latent variable $\mathbf{z}_i$ should read $\mathbf{Z}_i$.

Page 465

Paragraph –1, Line –1: If we adopt the representation (117), the factor $\widetilde{p}(\mathbf{X},\mathbf{Z}_j)$ should read $q_j^\star\left(\mathbf{Z}_j\right)$.

Page 466

Paragraph 1, Line 1: Again, $\widetilde{p}(\mathbf{X},\mathbf{Z}_j)$ should read $q_j^\star\left(\mathbf{Z}_j\right)$. The sentence "Thus we obtain..." should read, e.g., "Thus we see that we have already obtained a general expression for the optimal solution in (119)."

Page 467

The text before (10.12) and after (10.15): As Svensén and Bishop (2011) correct the left hand side of (10.12), we should write $q_1^\star(z_1)$ instead of $q^\star(z_1)$ and so on also in the text. Or, we should clarify that we simply write $q(\mathbf{z}_i)$ to denote the variational distribution over the latent variables $\mathbf{z}_i$ in the same manner as the notation for the probability $p(\cdot)$ is "overloaded" by its argument(s).

Page 468

The text after (10.16): The constant term in (10.16) is the *negative* entropy of $p(\mathbf{Z})$.

Page 470

The text after (10.19): "zero forcing" should be "zero-forcing" (with hyphenation).

Page 470

The text after (10.23): "Gaussian-Gamma" should read "Gaussian-gamma" (without capitalization for "gamma").

Page 478

Equation (10.63): The additive constant $+1$ on the right hand side should be omitted so that (10.63) should read

$$\nu_k = \nu_0 + N_k. \tag{120}$$

A quick check for the correctness of the re-estimation equations would be to consider the limit of $N \to 0$, in which the effective number of observations N_k also goes to zero and the re-estimation equations should reduce to identities. Equation (10.63) does not reduces to $\nu_k = \nu_0$, failing the test. Note that the solution for Exercise 10.13 given by Svensén and Bishop (2009) correctly derives the result (120).

Page 489

Equations (10.107) through (10.112): Some of the notations for the expectation are inconsistent with the one (1.36) employed in PRML; they should read $\mathbb{E}_{\mathbf{Z}}[\cdot]$ where $\mathbf{Z}$ is replaced with the corresponding latent variables. For example, $\mathbb{E}_{\alpha}\left[\ln q(\mathbf{w})\right]_{\mathbf{w}}$ in the last line of (10.107) and $\mathbb{E}\left[\ln p\left(\mathbf{t}|\mathbf{w}\right)\right]_{\mathbf{w}}$ in the left hand side of (10.108) should read $\mathbb{E}_{\mathbf{w}}\left[\ln q(\mathbf{w})\right]$ and $\mathbb{E}_{\mathbf{w}}\left[\ln p\left(\mathbf{t}|\mathbf{w}\right)\right]$, respectively, where we have assumed that the expectation $\mathbb{E}_{\mathbf{Z}}[\cdot]$ is taken with respect to the variational distribution $q\left(\mathbf{Z}\right)$.

Note however that we can safely omit the subscripts $\mathbf{Z}$ of the expectations $\mathbb{E}_{\mathbf{Z}}[\cdot]$ here, as we have done in, e.g., (10.70), because the variables over which we take the expectations are clear; we take the expectations over all the latent variables when we calculate the lower bound. We only need to make the subscripts explicit when we find an optimal factor $q^{\star}(\mathbf{Z}_i)$, in which case we take expectation selectively, that is, over all the latent variables but $\mathbf{Z}_i$; see, e.g., (10.92) and (10.96).

Page 490

Paragraph 3, Line 2: A comma (,) should be inserted after the ellipsis (. . .).

Page 496

Equation (10.140): The differential operator d should be an upright d. Moreover, the derivative of x with respect to x^2 should be written with parentheses as $\frac{\mathrm{d}x}{\mathrm{d}(x^2)}$, instead of $\frac{\mathrm{d}x}{\mathrm{d}x^2}$, to avoid ambiguity.

Page 499

Paragraph 1, Line –2: The sentence reads "Once [the right hand side of (10.152)] is normalized to give a variational posterior distribution $q(\mathbf{w})$, however, it no longer represents a bound." The statement does not make sense because the right hand side of (10.152) is a lower bound in terms of the variational parameters $\boldsymbol{\xi}$ and thus not directly dependent on the variational distribution $q(\mathbf{w})$. Moreover, as we shall see shortly, we obtain the optimal solution for $q(\mathbf{w})$ by making use of the general result (123) for local variational Bayes, but *not* by normalizing the right hand side of (10.152). Therefore, this sentence is irrelevant and can be safely removed.

Pages 500 and 501

Equations (10.156) and (10.160): It is not very clear why the variational posterior is obtained in the form (10.156) and the variational parameters can be optimized by maximizing (10.160). This EM-like algorithm is not the same as *the* EM algorithm we have seen in Chapter 9; it can be derived by maximizing the lower bound (10.3) as follows. Note that the discussion here is similar to, but more general than, that of Section 10.6.3.

In a more general setting, we consider a local variational approximation to the joint distribution of the form

$$p\left(\mathbf{X}, \mathbf{Z}\right) \geqslant \widetilde{p}\left(\mathbf{X}, \mathbf{Z}; \boldsymbol{\xi}\right) \tag{121}$$

where $\boldsymbol{\xi}$ denotes the set of variational parameters, assuming that we can bound the likelihood $p\left(\mathbf{X}|\mathbf{Z}\right) \geqslant \widetilde{p}\left(\mathbf{X}|\mathbf{Z}; \boldsymbol{\xi}\right)$ or the prior $p\left(\mathbf{Z}\right) \geqslant \widetilde{p}\left(\mathbf{Z}; \boldsymbol{\xi}\right)$, or both. Then, we can again bound the lower bound (10.3) as

$$\mathcal{L}(q) \geqslant \widetilde{\mathcal{L}}(q, \boldsymbol{\xi}) \equiv \mathbb{E}_{\mathbf{Z}}\left[\ln \widetilde{p}\left(\mathbf{X}, \mathbf{Z}; \boldsymbol{\xi}\right)\right] - \mathbb{E}_{\mathbf{Z}}\left[\ln q\left(\mathbf{Z}\right)\right] \tag{122}$$

where the expectation $\mathbb{E}_{\mathbf{Z}}[\cdot]$ is taken with respect to the variational distribution $q(\mathbf{Z})$. With much the same discussion as the derivation of the optimal solution (119) for the standard variational Bayesian method where we assume some appropriate factorization (10.5) for $q(\mathbf{Z})$, the optimal solution for the factor $q_j(\mathbf{Z}_j)$ that maximizes the lower bound $\widetilde{\mathcal{L}}(q, \boldsymbol{\xi})$ can be obtained by the relation

$$\ln q_j^{\star}\left(\mathbf{Z}_j\right) = \mathbb{E}_{\mathbf{Z} \setminus \mathbf{Z}_j}\left[\ln \widetilde{p}\left(\mathbf{X}, \mathbf{Z}; \boldsymbol{\xi}\right)\right] + \text{const} \tag{123}$$

which leads to the variational approximation to the posterior given by (10.156).

The optimization of the variational parameters $\boldsymbol{\xi}$ can be done by maximizing the first term of the lower bound $\widetilde{\mathcal{L}}(q, \boldsymbol{\xi})$, i.e.,

$$\mathcal{Q}(\boldsymbol{\xi}) = \mathbb{E}_{\mathbf{Z}}\left[\ln \widetilde{p}\left(\mathbf{X}, \mathbf{Z}; \boldsymbol{\xi}\right)\right] \tag{124}$$

which leads to the $\mathcal{Q}$ function given by (10.160).

Page 501

The text after (10.162): We have that the variational parameter $\lambda(\xi)$ is a monotonic function of ξ for $\xi \geqslant 0$, but not that its derivative $\lambda'(\xi)$ is.

Page 503

The text after (10.168): A period (.) should be appended at the end of the sentence that follows (10.168).

Page 504

Paragraph 1, Line 1: In order to obtain the optimized variational distribution (10.174), we should use the optimal solution (123) for *local* variational Bayes. Note that the result (123) is different from the result (119), or (10.9), for standard variational Bayes in that (123) is given in terms of the lower bound $\widetilde{p}\left(\mathbf{X}, \mathbf{Z}; \boldsymbol{\xi}\right)$ to the joint distribution $p\left(\mathbf{X}, \mathbf{Z}\right)$.

Page 512

Equation (10.222): The factor $(2\pi v_n)^{D/2}$ in the denominator of the right hand side should be omitted because it has been already included in the Gaussian in (10.213).

Page 513

Equations (10.223) and (10.224): The quantities v^{new} and $\mathbf{m}^{\text{new}}$ in (10.223) and (10.224) are different from those in (10.217) and (10.218).[8] Thus, we should introduce different notations, say, v and $\mathbf{m}$, with appropriate definitions. Specifically, one can rewrite the approximation to the model evidence in the form

$$p(\mathcal{D}) \simeq (2\pi v)^{D/2} \exp(B/2) \prod_{n=1}^{N} \left\{ s_n (2\pi v_n)^{-D/2} \right\} \tag{125}$$

where

$$B = \frac{\mathbf{m}^{\mathrm{T}}\mathbf{m}}{v} - \sum_{n=1}^{N} \frac{\mathbf{m}_n^{\mathrm{T}}\mathbf{m}_n}{v_n} \tag{126}$$

$$v^{-1} = \sum_{n=1}^{N} v_n^{-1} \tag{127}$$

$$v^{-1}\mathbf{m} = \sum_{n=1}^{N} v_n^{-1}\mathbf{m}_n. \tag{128}$$

Page 515

Equations (10.228) and (10.229): Although Svensén and Bishop (2011) correct (10.228) so that $q^{\backslash b}(\mathbf{x})$ is a normalized distribution, we do not need the normalization of $q^{\backslash b}(\mathbf{x})$ here and, even with this normalization, we cannot ensure that $\widehat{p}(\mathbf{x})$ given by (10.229) is normalized. Similarly to (10.195), we can proceed with the unnormalized $q^{\backslash b}(\mathbf{x})$ given by the original (10.228) and, rather than correcting (10.228), we should correct (10.229) so that

$$\widehat{p}(\mathbf{x}) \propto q^{\backslash b}(\mathbf{x}) f_b(x_2, x_3) = \ldots \tag{129}$$

implying that $\widehat{p}(\mathbf{x})$ is a normalized distribution.

Page 515

The text after (10.229): The new distribution $q^{\text{new}}(\mathbf{z})$ should read $q^{\text{new}}(\mathbf{x})$.

Page 516

Equation (10.240): The subscript k of the product $\prod_k \ldots$ should read $k \neq j$ because we have already removed the term $\widetilde{f}_j(\boldsymbol{\theta}_j)$.

[8] See Svensén and Bishop (2011) for the errata for (10.217) and (10.218).

Pages 554 and 555

Equation (11.72), Line –2 and the text after (11.72): The expectation in the last line but one of (11.72) is taken with respect to the probability $p_G(\mathbf{z})$. This is probably better expressed in words, rather than the unclear notation like $\mathbb{E}_{G(\mathbf{z})}[\cdot]$. Specifically, the expectation should read

$$\mathbb{E}_{\mathbf{z}}\left[\exp\left(-E(\mathbf{z})+G(\mathbf{z})\right)\right] \tag{130}$$

where we have written the argument $\mathbf{z}$ for $E(\mathbf{z})$ and $G(\mathbf{z})$ for clarity; and the text following (11.72) should read "where $\mathbb{E}_{\mathbf{z}}[\cdot]$ is taken with respect to $p_G(\mathbf{z})$ and $\{\mathbf{z}^{(l)}\}$ are samples drawn from the distribution defined by $p_G(\mathbf{z})$."

Page 556

Exercise 11.7, Line 1: The interval should be $[-\pi/2, \pi/2]$ instead of $[0, 1]$.

Page 557

Exercise 11.14, Line 2: The variance should be σ_i^2 instead of σ_i.

Page 564

The text after (12.12): The derivative we consider here is that with respect to b_j (*not* that with respect to b_i).

Page 564

Paragraph –1, Line 2: $\mathbf{u}_i = 0$ should read $\mathbf{u}_i = \mathbf{0}$ (the right hand side should be a zero vector $\mathbf{0}$).

Page 575

Paragraph –2, Line 5: The zero vector should be a row vector instead of a column vector so that we have $\mathrm{v}^{\mathrm{T}}\mathbf{U} = \mathbf{0}^{\mathrm{T}}$. Or, the both sides are transposed to give $\mathbf{U}^{\mathrm{T}}\mathbf{v} = \mathbf{0}$.

Page 578

Equation (12.53): As stated in the text preceding (12.53), we should substitute $\boldsymbol{\mu} = \overline{\mathbf{x}}$ into (12.53).

Page 578

The text before (12.56): For the maximization with respect to $\mathbf{W}$, we use (C.25) and (C.27) instead of (C.24).

Page 579

Paragraph 1, Line 5: The eigendecomposition requires $O(D^3)$ computations (in the plural form).

Page 599

Exercise 12.1, Line –1: The quantity λ_{M+1} is an eigenvalue (not an eigenvector).

Page 602

Exercise 12.25, Line 2: The latent space distribution should read $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$.

Page 610

Paragraph 1, Line –4: The text "our predictions for $\mathbf{x}_{n+1}$ depends on..." should read: "our predictions for $\mathbf{x}_{n+1}$ depend on..." (Remove the trailing 's' from the verb).

Page 620

Paragraph –1, Line 4 and the following (unlabeled) equation: The last sentence before the equation and the equation should each be terminated with a period (.).

Pages 621 and 622

Figures 13.12 and 13.13: It should be clarified that, similarly to the notations $\alpha(z_{nk})$ and $\beta(z_{nk})$, $p(\mathbf{x}_n|z_{nk})$ denotes the value of $p(\mathbf{x}_n|\mathbf{z}_n)$ when $z_{nk} = 1$.

Page 622

The text after (13.39): "we see" should be omitted.

Page 622

Equation (13.40): The summations should read $\sum_{n=1}^{N}$.

Page 623

Paragraph 1, Line –2: z_{nk} should read $z_{n-1,k}$.

Page 631

Equation (13.73): The equation should read

$$\sum_{r=1}^{R} \ln \left\{ \frac{p(\mathbf{X}_r|\boldsymbol{\theta}_{m_r})\, p(m_r)}{\sum_{l=1}^{M} p(\mathbf{X}_r|\boldsymbol{\theta}_l)\, p(l)} \right\}. \tag{131}$$

Page 637

Equations (13.81), (13.82), and (13.83): The distribution (13.81) over $\mathbf{w}$ should read

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\Gamma}) \tag{132}$$

and so on.

Page 638

Paragraph 1, Line 2: "conditional on" should read "conditioned on."

Page 641

Equation (13.104) and the preceding text: The form of the Gaussian is unclear. Since a multivariate Gaussian is usually defined over a column vector, we should construct a column vector from the concerned random variables to clearly define the mean and the covariance. Specifically, (13.104) and the preceding text should read: ... we see that $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ is a Gaussian of the form

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = \mathcal{N}\left(\begin{pmatrix}\mathbf{z}_{n-1}\\ \mathbf{z}_n\end{pmatrix}\middle|\begin{pmatrix}\widehat{\boldsymbol{\mu}}_{n-1}\\ \widehat{\boldsymbol{\mu}}_n\end{pmatrix}, \begin{pmatrix}\widehat{\mathbf{V}}_{n-1} & \widehat{\mathbf{V}}_{n-1,n}\\ \widehat{\mathbf{V}}_{n-1,n}^{\mathrm{T}} & \widehat{\mathbf{V}}_n\end{pmatrix}\right) \tag{133}$$

where the mean $\widehat{\boldsymbol{\mu}}_n$ and the covariance $\widehat{\mathbf{V}}_n$ of $\mathbf{z}_n$ are given by (13.100) and (13.101), respectively; and the covariance $\widehat{\mathbf{V}}_{n-1,n}$ between $\mathbf{z}_{n-1}$ and $\mathbf{z}_n$ is given by

$$\widehat{\mathbf{V}}_{n-1,n} = \operatorname{cov}\left[\mathbf{z}_{n-1}, \mathbf{z}_n\right] = \mathbf{J}_{n-1}\widehat{\mathbf{V}}_n. \tag{134}$$

Pages 642 and 643

Equation (13.109) and the following equations: If we follow the notation in Chapter 9, the typeface of the Q function should be $\mathcal{Q}$.

Page 642

Equation (13.109): If we follow the notation for a conditional expectation given by (1.37), (13.109) should read

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \mathbb{E}_{\mathbf{Z}}\left[\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}\right)\middle|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right] \tag{135}$$

$$= \int \mathrm{d}\mathbf{Z}\; p\left(\mathbf{Z}\middle|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right) \ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}\right) \tag{136}$$

which corresponds to (9.30).

Page 643

Equation (13.111): $\mathbf{V}_0^{\text{new}}$ should read $\mathbf{P}_0^{\text{new}}$. Svensén and Bishop (2011) have failed to mention (13.111).

Page 643

Equation (13.114): The size of the opening curly brace "{" should match that of the closing curly brace "}."

Page 647

The caption of Figure 13.23, Line –1: $p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}^{(l)})$ should read $p(\mathbf{x}_{n+1}|z_{n+1}^{(l)})$.

Page 649

Exercise 13.14, Line 1: (8.67) should be (8.64).

Page 650

Equations (13.127) and (13.128): The equal signs should be aligned.

Page 651

Exercises 13.25 through 13.28: A zero matrix is denoted by $\mathbf{O}$ (*not* by $\mathbf{0}$ nor 0) so that we should write $\mathbf{A} = \mathbf{O}$ and so on.

Page 658

The equation at the bottom of Figure 14.1: The subscript of the summation in the right hand side should read $m = 1$.

Page 668

Equation (14.37): The arguments of the probability are notationally inconsistent with those of (14.34), (14.35), and (14.36). Specifically, the conditioning on $\boldsymbol{\phi}_n$ should read that on t_n and the probability $p(k|\dots)$ be the value of $p(\mathbf{z}_n|\dots)$ when $z_{nk} = 1$, which we write $p(z_{nk} = 1|\dots)$. Moreover, strictly speaking, the old parameters $\pi_k, \mathbf{w}_k, \beta$ should read $\pi_k^{\text{old}}, \mathbf{w}_k^{\text{old}}, \beta^{\text{old}} \in \boldsymbol{\theta}^{\text{old}}$. In order to solve these problems, we should rewrite (14.37) as, for example,

$$\gamma_{nk} = \mathbb{E}\left[z_{nk}\middle|t_n, \boldsymbol{\theta}^{\text{old}}\right] \tag{137}$$

where we have written the conditioning in the expectation explicitly and the expectation is given by

$$\mathbb{E}\left[z_{nk}|t_n, \boldsymbol{\theta}\right] = p\left(z_{nk} = 1|t_n, \boldsymbol{\theta}\right) = \frac{\pi_k \mathcal{N}\left(t_n\middle|\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}\right)}{\sum_j \pi_j \mathcal{N}\left(t_n\middle|\mathbf{w}_j^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}\right)}. \tag{138}$$

Page 668

The unlabeled equation between (14.37) and (14.38): If we write the implicit conditioning in the expectation explicitly (similarly to the above equations), the unlabeled equation should read

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{t}, \mathbf{Z}|\boldsymbol{\theta})\middle|\mathbf{t}, \boldsymbol{\theta}^{\text{old}}\right] \tag{139}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}\left\{\ln \pi_k + \ln \mathcal{N}\left(t_n\middle|\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}\right)\right\}. \tag{140}$$

Page 669

Equations (14.40) and (14.41): The left hand sides should both read a zero vector $\mathbf{0}$ instead of a scalar zero 0.

Page 669

Equation (14.41): $\mathbf{\Phi}$ is undefined. The text following (14.41) should read for example: where $\mathbf{R}_k = \mathrm{diag}(\gamma_{nk})$ is a diagonal matrix of size $N \times N$ and $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N)^{\mathrm{T}}$ is an $N \times M$ matrix. Here, N is the size of the data set and M is the dimensionality of the feature vectors $\boldsymbol{\phi}_n$.

Page 669

Equation (14.43): "+const" should be added to the right hand side.

Page 671

The text after (14.46): The text should read: "where we have omitted the dependence on $\{\boldsymbol{\phi}_n\}$ and defined $y_{nk} = \ldots$" Or, $\boldsymbol{\phi}$ should have been omitted from the left hand side of (14.45) in the first place.

Page 671

Equation (14.48): The notation should be corrected similarly to (137) and (138).

Page 671

Equation (14.49): The notation should be corrected similarly to (139).

Page 672

Equation (14.52): The negation should be removed so that the Hessian is given by $\mathbf{H}_k \equiv \nabla_k \nabla_k \mathcal{Q}$ where

$$\nabla_k \nabla_k \mathcal{Q} = -\sum_{n=1}^{N} \gamma_{nk} y_{nk} (1 - y_{nk}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}. \tag{141}$$

Page 674

Exercise 14.1, Line 1: "of" should be inserted after "set."

Page 686

Equation (B.9): The mode (B.9) of the beta distribution exists "if $a > 1$ and $b > 1$."

Page 686

Paragraph –1, Line –3: The comma in the first inline math should be removed so that the product reads: $m \times (m-1) \times \cdots \times 2 \times 1$.

Page 687

Equation (B.20): The mode (B.20) of the Dirichlet exists "if $\alpha_k > 1$ for all k."

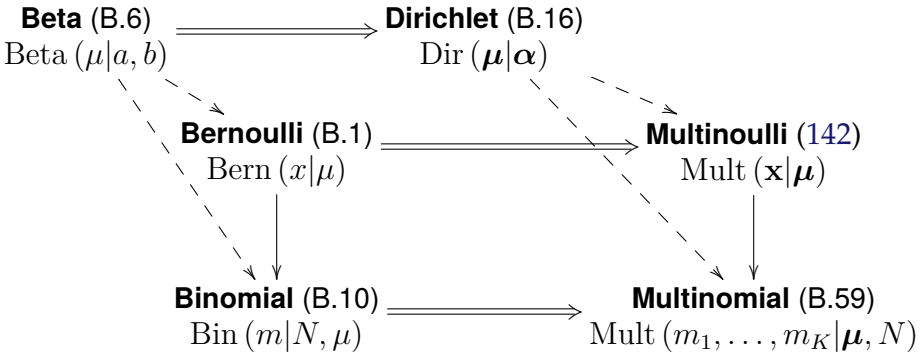

Figure 4 The relationship between discrete distributions and their conjugate priors. Here, A $\Rightarrow$ B denotes "A generalizes to B with multiple categories concerned"; and A $\rightarrow$ B "A to B with multiple observations." Note also that A $\dashrightarrow$ B denotes "A is the conjugate prior for B."

Page 687

Equation (B.25): The differential operator d should be an upright d.

Page 688

Paragraph 1, Line 1: "Gamma" should read "gamma" (without capitalization).

Page 689

Paragraph 1, Line 1: "positive-definite" should read "positive definite" (without hyphenation).

Page 689

Equation (B.49): $\mathbf{x}$ in the right hand side should read $\mathbf{x}_a$.

Page 690

Equation (B.52): μ_o in the right hand side should read μ_0 (the subscript should be a zero 0).

Page 690

Equation (B.54): The discrete distribution of the form (B.54), or (2.26), is known as the *categorical* or the *multinoulli* distribution (Murphy, 2012). It is also sometimes called, less precisely, the "multinomial" or the "discrete" distribution. Of these terms, I would prefer the term *multinoulli* because it naturally suggests that it is a generalization of the *Bernoulli* distribution (B.1) to multiple categories $K > 2$ and also a special case of the *multinomial* distribution (B.59) where we have only a single observation $N = 1$ (see Figure 4 for the relationship between the discrete distributions found in PRML). Since we often make use of this discrete distribution, we shall introduce some notation for the right hand side of (B.54).

Multinoulli The *multinoulli* distribution is a distribution over the K-dimensional binary variable $\mathbf{x} = (x_1, \ldots, x_K)^{\mathrm{T}}$ where $x_k \in \{0, 1\}$ such that $\sum_k x_k = 1$, i.e., we employ the one-of-K coding scheme for $\mathbf{x}$. Here, we "overload" the notation (B.59) for the multinomial and write the multinoulli as

$$\operatorname{Mult}(\mathbf{x}|\boldsymbol{\mu}) \equiv \prod_{k=1}^{K} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{K} x_k \ln \mu_k\right\} \tag{142}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$ are probabilities such that $0 \leqslant \mu_k \leqslant 1$ and $\sum_k \mu_k = 1$.

When we identify a multinoulli distribution from its functional form, e.g., in the posterior distribution (113) for the Gaussian mixture model (114) of Section 9.2, one will find it helpful to know that the multinoulli distribution (142) can also be expressed in terms of unnormalized probabilities $\widetilde{\mu}_k \geqslant 0$, i.e.,

$$\operatorname{Mult}(\mathbf{x}|\boldsymbol{\mu}) \propto \prod_{k=1}^{K} \widetilde{\mu}_k^{x_k} = \exp\left\{\sum_{k=1}^{K} x_k \ln \widetilde{\mu}_k\right\} \tag{143}$$

where the normalized probabilities μ_k can be found

$$\mu_k = p(x_k = 1) = \frac{\widetilde{\mu}_k}{\sum_j \widetilde{\mu}_j}. \tag{144}$$

Page 691

The icon for Student's t-distribution: As we have seen in the erratum for Figure 2.15, the tails of the t-distributions are too high. Figure 3 gives the correct plot.

Page 692

Equation (B.68): This form of multivariate Student's t-distribution is derived in Section 2.3.7 by marginalizing over the gamma distributed (scalar) variable η in (2.161), but *not* by marginalizing over the $D \times D$ precision matrix $\boldsymbol{\Lambda}$ that is governed by the Wishart distribution $\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu)$ where $\mathbf{W} \succ 0$ and $\nu > D - 1$, which results in a marginal distribution of the form

$$p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{W}, \nu) = \int \mathcal{N}\left(\mathbf{x}\middle|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu)\, \mathrm{d}\boldsymbol{\Lambda}. \tag{145}$$

The above marginal distribution (145) is indeed equivalent to (B.68) with some reparameterization. However, this result is not so obvious that I would like to show it here. Note that such marginalization is also used to derive a mixture of Student's t-distributions given by (10.81) in Exercise 10.19. The key idea is that the integrand in the right hand side of (145) can be identified as an unnormalized Wishart distribution and the marginalization can be done in a symbolic manner. More specifically, we have

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{W}, \nu) = \int \mathrm{d}\boldsymbol{\Lambda}\, & \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})\right\} \\ & \times B(\mathbf{W}, \nu) |\boldsymbol{\Lambda}|^{(\nu - D - 1)/2} \exp\left\{-\frac{1}{2} \operatorname{Tr}\left(\mathbf{W}^{-1}\boldsymbol{\Lambda}\right)\right\} \qquad (146) \\ = & \frac{2^{(\nu+1)D/2} \Gamma_D\left(\frac{\nu+1}{2}\right) \left|\mathbf{W}^{-1} + (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\right|^{-(\nu+1)/2}}{(2\pi)^{D/2} 2^{\nu D/2} \Gamma_D\left(\frac{\nu}{2}\right) |\mathbf{W}|^{\nu/2}} \qquad (147) \end{aligned}$$

where we have used the multivariate gamma function (Anderson, 2003; Olver et al., 2016) given by[9]

$$\Gamma_D\left(\frac{\nu}{2}\right) = \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu+1-i}{2}\right) \tag{149}$$

so that we can write the normalization constant (B.79) as

$$B(\mathbf{W},\nu)^{-1} = 2^{\nu D/2} \left|\mathbf{W}\right|^{\nu/2} \Gamma_D\left(\frac{\nu}{2}\right). \tag{150}$$

Finally, we obtain

$$p\left(\mathbf{x}|\boldsymbol{\mu},\mathbf{W},\nu\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu+1-D}{2}\right)} \frac{\left|\mathbf{W}\right|^{1/2}}{\pi^{D/2}} \left[1 + \left(\mathbf{x}-\boldsymbol{\mu}\right)^{\mathrm{T}} \mathbf{W} \left(\mathbf{x}-\boldsymbol{\mu}\right)\right]^{-(\nu+1)/2} \tag{151}$$

where we have used (149) and (C.15). Thus, we see that the marginal distribution of the form (145) is equivalent to the multivariate Student's t-distribution of the form (B.68) or (2.162); they are related by

$$p\left(\mathbf{x}|\boldsymbol{\mu},\mathbf{W},\nu\right) = \mathrm{St}\left(\mathbf{x}|\boldsymbol{\mu},(\nu+1-D)\mathbf{W},\nu+1-D\right). \tag{152}$$

If the scale matrix is isotropic, which is common in practice, so that $\mathbf{W} = \widetilde{W}\mathbf{I}$ where $\widetilde{W} > 0$, then the resulting multivariate Student's t-distribution (152) is again isotropic. The same marginal distribution can also be obtained by marginalizing with respect to a univariate Wishart (gamma) prior so that

$$\int_0^\infty \mathcal{N}\left(\mathbf{x}\middle|\boldsymbol{\mu},\widetilde{\lambda}^{-1}\mathbf{I}\right) \mathcal{W}\left(\widetilde{\lambda}\middle|\widetilde{W},\widetilde{\nu}\right) \mathrm{d}\widetilde{\lambda} = \mathrm{St}\left(\mathbf{x}\middle|\boldsymbol{\mu},\widetilde{\nu}\widetilde{W}\mathbf{I},\widetilde{\nu}\right) \tag{153}$$

where $\widetilde{\nu} = \nu + 1 - D > 0$. Note that the "covariance" parameter (44) of the corresponding multivariate Wishart prior $\mathcal{W}\left(\boldsymbol{\Lambda}|\mathbf{W},\nu\right)$ for which we obtain the same marginal (153) is however *not* equal to $\widetilde{\sigma}^2\mathbf{I}$ where $\widetilde{\sigma}^2 = \left(\widetilde{\nu}\widetilde{W}\right)^{-1}$ is the "covariance" parameter of the univariate Wishart prior $\mathcal{W}\left(\widetilde{\lambda}\middle|\widetilde{W},\widetilde{\nu}\right)$, but is given by

$$\boldsymbol{\Sigma} = \left(\nu\mathbf{W}\right)^{-1} = \frac{\widetilde{\nu}}{\widetilde{\nu}-1+D}\widetilde{\sigma}^2\mathbf{I}. \tag{154}$$

So far, we have observed that a marginal distribution of the form (145) where the marginalization is taken over a matrix-valued random variable $\boldsymbol{\Lambda}$ is equivalent to a marginal distribution of the form (2.161) or, if the scale matrix is isotropic, of the form (153) where the marginalization is over a scalar random variable η or $\widetilde{\lambda}$, respectively. Given that those marginals reduce to an identical multivariate Student's t-distribution (with some reparameterization), we now have a natural question: *Which form of marginal is better than the*

[9]The multivariate gamma function $\Gamma_D(\cdot)$ is defined by

$$\Gamma_D(a) \equiv \int_{\mathbf{X}\succ 0} \left|\mathbf{X}\right|^{a-(D+1)/2} \exp(-\operatorname{Tr}(\mathbf{X}))\,\mathrm{d}\mathbf{X} \tag{148}$$

where the integration is taken over the space of symmetric positive-definite matrices (Anderson, 2003; Olver et al., 2016). One can see that, when $D = 1$, the multivariate gamma function $\Gamma_D(\cdot)$ reduces to the (univariate) gamma function $\Gamma(\cdot)$ defined by (1.141).

other? I would argue that a marginal with fewer latent variables, i.e., (2.161) or (153), is always better than a marginal with more latent variables, i.e., (145), because fewer latent variables imply less computational space and complexity as well as a tighter bound on the (marginal) likelihood and thus faster convergence when we infer a model involving such marginals with the EM algorithm (see Chapter 9) or variational methods (Chapter 10). Moreover, the marginal of the form (2.161) enjoys even greater modeling flexibility in that it allows us to learn the mean $\boldsymbol{\mu}$ and the precision $\boldsymbol{\Lambda}$ parameters with, e.g., maximum likelihood (see Exercise 12.24) or variational Bayes by introducing a (conditionally) conjugate prior for $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ (Svensén and Bishop, 2005).

Page 693

Equations (B.78) through (B.82): Some appropriate citation, e.g., Anderson (2003), is needed for the Wishart distribution because it has been introduced in Section 2.3.6 without any proof for the normalization constant as well as other statistics.

Page 693

Paragraph –1, Line –4: "Gamma" should read "gamma" (without capitalization).

Page 693

Paragraph –1, Line –1: $b = 1/2W$ should read $b = 1/(2W)$ for clarity.

Page 696

Equation (C.5): Replacing $\mathbf{B}^{\mathrm{T}}$ with $\mathbf{A}$, we obtain a more general identity

$$\left(\mathbf{P}^{-1} + \mathbf{A}\mathbf{R}^{-1}\mathbf{B}\right)^{-1}\mathbf{A}\mathbf{R}^{-1} = \mathbf{P}\mathbf{A}\left(\mathbf{B}\mathbf{P}\mathbf{A} + \mathbf{R}\right)^{-1}. \tag{155}$$

The identity (155) is necessary to show the *push-through identity* (C.6), which in turn can be used to show *Sylvester's determinant identity* (C.14). As suggested in the text, the above identity (155) can be directly verified by right multiplying both sides by $(\mathbf{B}\mathbf{P}\mathbf{A} + \mathbf{R})$. However, I would prefer to prove the general push-through identity (155) together with the *Woodbury identity* (C.7) in terms of the inverse of a partitioned matrix, which we have already seen in Section 2.3.1. To this end, we first introduce a square matrix $\mathbf{M}$ that is partitioned into four submatrices so that

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \tag{156}$$

where $\mathbf{A}$ and $\mathbf{D}$ are square (but not necessarily the same dimension) and then note that $\mathbf{M}$ can be block diagonalized as

$$\begin{pmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{I} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{M}/\mathbf{A} \end{pmatrix} \tag{157}$$

or

$$\begin{pmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{M}/\mathbf{D} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{pmatrix} \tag{158}$$

if $\mathbf{A}$ or $\mathbf{D}$ is nonsingular, respectively, where we have written the *Schur complement* of $\mathbf{M}$ with respect to $\mathbf{A}$ or $\mathbf{D}$ as

$$\mathbf{M}/\mathbf{A} \equiv \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \tag{159}$$

or

$$\mathbf{M}/\mathbf{D} \equiv \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \tag{160}$$

respectively.[10] The above block diagonalization identities (157) and (158) yield two versions of the inverse partitioned matrix $\mathbf{M}^{-1}$, i.e.,

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{I} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{O} \\ \mathbf{O} & (\mathbf{M}/\mathbf{A})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \tag{163}$$

$$= \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{M}/\mathbf{A}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{M}/\mathbf{A}\right)^{-1} \\ -\left(\mathbf{M}/\mathbf{A}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & \left(\mathbf{M}/\mathbf{A}\right)^{-1} \end{pmatrix} \tag{164}$$

and

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{pmatrix} \begin{pmatrix} (\mathbf{M}/\mathbf{D})^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} \tag{165}$$

$$= \begin{pmatrix} (\mathbf{M}/\mathbf{D})^{-1} & -\left(\mathbf{M}/\mathbf{D}\right)^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{M}/\mathbf{D}\right)^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\left(\mathbf{M}/\mathbf{D}\right)^{-1}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \tag{166}$$

respectively. Equating the right hand sides, we have, e.g.,

$$(\mathbf{M}/\mathbf{D})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{M}/\mathbf{A}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{167}$$

and

$$-\left(\mathbf{M}/\mathbf{A}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} = -\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{M}/\mathbf{D}\right)^{-1}. \tag{168}$$

Substituting (159) and (160) into both sides and replacing $\mathbf{D}$ with $-\mathbf{D}$, we finally have

$$\left(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{169}$$

and

$$\left(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} = \mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} \tag{170}$$

which are equivalent to (C.7) and (155), respectively.

Page 696

Paragraph 3, Line 2: $\sum_n \alpha_n \mathbf{a}_n = 0$ should read $\sum_n \alpha_n \mathbf{a}_n = \mathbf{0}$ (the right hand side should be a zero vector $\mathbf{0}$).

[10] Note that the notation for the Schur complement is chosen to suggest that it has a flavor of division (Minka, 2000). In fact, taking the determinant on both sides of (157) and (158), we have from the *Leibniz formula for determinants* (C.10) that

$$\det(\mathbf{M}) = \det(\mathbf{A})\det(\mathbf{M}/\mathbf{A}) \tag{161}$$

and

$$\det(\mathbf{M}) = \det(\mathbf{D})\det(\mathbf{M}/\mathbf{D}) \tag{162}$$

respectively.

Pages 696 and 697

Equations (C.8), (C.9), and (C.12): Note that, although matrix trace $\mathrm{Tr}(\cdot)$ only applies to square matrices, the matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ in (C.8) and (C.9) themselves are not necessarily square. On the other hand, in order for the determinant identity (C.12) to hold, both $\mathbf{A}$ and $\mathbf{B}$ must be square.

Page 697

Equation (C.17): It is clear that the definition (C.17) of the derivative of a scalar with respect to a vector and that (C.18) of the derivative of a vector with respect to a vector contradict each other. The vector derivative of the form (C.17) is usually called the *gradient* whereas (C.18) is called the *Jacobian* (Minka, 2000). Note that (C.16) is a special case of (C.18) and thus the Jacobian. We should use a different notation, say, ∇ for the gradient to avoid ambiguity. More specifically, given a vector function $\mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}), \ldots, y_M(\mathbf{x}))^{\mathrm{T}}$ where $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$, we write the gradient of $\mathbf{y}(\mathbf{x})$ with respect to $\mathbf{x}$ as

$$\nabla_{\mathbf{x}}\mathbf{y} \equiv \left(\frac{\partial y_j}{\partial x_i}\right) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_M}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_D} & \cdots & \frac{\partial y_M}{\partial x_D} \end{pmatrix}. \tag{171}$$

As a special case, we see that the gradient of a scalar function $y(\mathbf{x})$ with respect to a column vector $\mathbf{x}$ is again a column vector of the same dimension, corresponding to the right hand side of (C.17), i.e.,

$$\nabla_{\mathbf{x}} y = \left(\frac{\partial y}{\partial x_i}\right) = \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_D} \end{pmatrix}. \tag{172}$$

Note also that the right hand side of the definition of the gradient (171) is identical to the transpose of the Jacobian matrix $\partial\mathbf{y}/\partial\mathbf{x} = (\partial y_i/\partial x_j)$ so that $\nabla_{\mathbf{x}}\mathbf{y} = (\partial\mathbf{y}/\partial\mathbf{x})^{\mathrm{T}}$, as a consequence of which the chain rule for the gradient is such that the intermediate gradients are built up "towards the left," i.e.,

$$\nabla_{\mathbf{x}}\mathbf{z}(\mathbf{y}) = \left(\frac{\partial\mathbf{z}}{\partial\mathbf{y}}\frac{\partial\mathbf{y}}{\partial\mathbf{x}}\right)^{\mathrm{T}} = \nabla_{\mathbf{x}}\mathbf{y}\nabla_{\mathbf{y}}\mathbf{z}. \tag{173}$$

Since the chain rule (173) is handy when we compute the gradients of composite functions (see below), I would suggest that it should also be pointed out in the "(Vector and) Matrix Derivatives" section of Appendix C.

At this point, one might wonder why we use the two different forms of vector derivative that are identical up to the transposed layout, i.e., the gradient $\nabla_{\mathbf{x}}\mathbf{y}$ and the Jacobian $\partial\mathbf{y}/\partial\mathbf{x}$. As Minka (2000) points out, Jacobians are useful in calculus while gradients are useful in optimization. For instance, we can write down the Taylor series expansion (up to the second order) of a scalar function $f(\mathbf{x})$ succinctly in terms of the gradients as

$$f(\mathbf{x} + \epsilon\boldsymbol{\eta}) = f(\mathbf{x}) + \epsilon\boldsymbol{\eta}^{\mathrm{T}}\mathbf{g}(\mathbf{x}) + \frac{\epsilon^2}{2}\boldsymbol{\eta}^{\mathrm{T}}\mathbf{H}(\mathbf{x})\boldsymbol{\eta} + O\left(\epsilon^3\right) \tag{174}$$

where $\mathbf{g}(\mathbf{x})$ and $\mathbf{H}(\mathbf{x})$ are the gradient vector and the Hessian matrix of $f(\mathbf{x})$, respectively, so that

$$\mathbf{g}(\mathbf{x}) \equiv \nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_D} \end{pmatrix}, \qquad \mathbf{H}(\mathbf{x}) \equiv \nabla\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_D \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_D \partial x_D} \end{pmatrix}. \tag{175}$$

Page 697

Equation (C.19): Following the gradient notation (171), (C.19) should read

$$\nabla \left\{ \mathbf{x}^{\mathrm{T}} \mathbf{a} \right\} = \nabla \left\{ \mathbf{a}^{\mathrm{T}} \mathbf{x} \right\} = \mathbf{a} \tag{176}$$

where we have omitted the subscript $\mathbf{x}$ in what should be $\nabla_{\mathbf{x}}$. Some other useful identities I would suggest to include are

$$\nabla \left\{ \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x} \right\} = \nabla \operatorname{Tr} \left(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mathbf{A} \right) = \left(\mathbf{A}^{\mathrm{T}} + \mathbf{A} \right) \mathbf{x} \tag{177}$$

$$\nabla \left\{ \mathbf{B}\mathbf{x} \right\} = \mathbf{B}^{\mathrm{T}} \tag{178}$$

$$\nabla \left\{ \phi \mathbf{y} \right\} = \nabla \phi \mathbf{y}^{\mathrm{T}} + \phi \nabla \mathbf{y} \tag{179}$$

where matrices $\mathbf{A}$ and $\mathbf{B}$ are constants. Note that $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ in (177) is a quadratic form and thus the square matrix $\mathbf{A}$ is usually taken to be symmetric so that $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$, in which case we have

$$\nabla \left\{ \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x} \right\} = 2\mathbf{A}\mathbf{x}. \tag{180}$$

Substituting $\mathbf{A} = \mathbf{I}$ gives

$$\nabla \left\| \mathbf{x} \right\|^2 = 2\mathbf{x} \tag{181}$$

where $\left\| \mathbf{x} \right\| = \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}}$ is the norm of $\mathbf{x}$. We make use of the above identity (181) when, e.g., we take the gradient of a sum-of-squares error function of the form (3.12), which can be expressed in terms of the design matrix $\mathbf{\Phi}$ given by (3.16) as

$$E(\mathbf{w}) = \frac{1}{2} \left\| \mathbf{t} - \mathbf{\Phi}\mathbf{w} \right\|^2. \tag{182}$$

Taking the gradient of (182) with respect to $\mathbf{w}$, we have

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = -\mathbf{\Phi}^{\mathrm{T}} \left(\mathbf{t} - \mathbf{\Phi}\mathbf{w} \right) \tag{183}$$

where we have used the identity (181) together with the chain rule (173) and the identity (178). The same result can also be obtained by first expanding the square norm in (182) and then differentiating it using the gradient identities given above. We use (179) when, e.g., we evaluate the Hessian (5.83) of a nonlinear sum-of-squares error function such as (5.82), which takes the form

$$J = \frac{1}{2} \sum_{n=1}^{N} \varepsilon_n^2 = \frac{1}{2} \left\| \boldsymbol{\varepsilon} \right\|^2 \tag{184}$$

where we have written $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^{\mathrm{T}}$. The gradient and the Hessian of J are evaluated as

$$\nabla J = \sum_{n=1}^{N} \varepsilon_n \nabla \varepsilon_n = (\nabla \boldsymbol{\varepsilon})\, \boldsymbol{\varepsilon} \tag{185}$$

$$\nabla\nabla J = \sum_{n=1}^{N} \nabla \varepsilon_n \left(\nabla \varepsilon_n\right)^{\mathrm{T}} + \sum_{n=1}^{N} \varepsilon_n \nabla\nabla \varepsilon_n \tag{186}$$

$$= \nabla \boldsymbol{\varepsilon} \left(\nabla \boldsymbol{\varepsilon}\right)^{\mathrm{T}} + \sum_{n=1}^{N} \varepsilon_n \nabla\nabla \varepsilon_n \tag{187}$$

respectively. The second form of the Hessian, which, however, does not necessarily result in efficient implementation (neither does that of the gradient), can be directly obtained by using the identity

$$\nabla \left\{\mathbf{R}\boldsymbol{\phi}\right\} = \nabla \boldsymbol{\phi} \mathbf{R}^{\mathrm{T}} + \sum_{m=1}^{M} \phi_m \nabla \mathbf{r}_m \tag{188}$$

where $\mathbf{R} = (\mathbf{r}_1, \ldots, \mathbf{r}_M)$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_M)^{\mathrm{T}}$. One can see that (178) and (179) are special cases of (188).

Page 698

Equation (C.20): Although the Jacobian of a vector with respect to a vector is defined in (C.18), the Jacobian of a matrix with respect to a scalar has not been defined. The Jacobian $\partial \mathbf{A}/\partial x$ of a matrix $\mathbf{A} = (A_{ij})$ with respect to a scalar x can be defined as a matrix with the same dimensionality as $\mathbf{A}$ so that

$$\frac{\partial \mathbf{A}}{\partial x} \equiv \left(\frac{\partial A_{ij}}{\partial x}\right) \tag{189}$$

which is analogous to (C.18) in that the partial derivatives are laid out according to the numerator, i.e., $\mathbf{A}$. On the other hand, the gradient (171) is such that the derivatives are laid out according to the denominator. In a similar analogy, we can define the gradient $\nabla_{\mathbf{A}} y$ of a scalar y with respect to a matrix $\mathbf{A}$ as

$$\nabla_{\mathbf{A}} y \equiv \left(\frac{\partial y}{\partial A_{ij}}\right). \tag{190}$$

Page 698

Equation (C.22): For this identity to be well-defined, it is necessary that we have $\det(\mathbf{A}) > 0$. We should make this assumption clear. Or, if we adopt the absolute determinant notation (15) for $|\mathbf{A}|$, the identity (C.22) holds, in fact, for any nonsingular $\mathbf{A}$ such that $\det(\mathbf{A}) \neq 0$ as we shall see shortly.

The section named "Eigenvector Equation" of Appendix C gives us a hint for a proof of (C.22) where $\mathbf{A}$ is assumed to be symmetric positive definite so that $\mathbf{A} \succ 0$. Although the restricted proof outlined in PRML is indeed highly instructive, we need a more general proof because we make use of this identity, e.g., in Exercise 2.34 without the assumptions required by the restricted proof.[11]

[11]Note that one can easily extend the restricted proof of (C.22) for a symmetric positive-definite matrix $\mathbf{A}$

Jacobi's formula To this end, we first show *Jacobi's formula*, which is an identity that holds for any square matrix $\mathbf{A}$ given by

$$\frac{\partial}{\partial x}\det(\mathbf{A}) = \mathrm{Tr}\left(\mathbf{A}^\dagger \frac{\partial \mathbf{A}}{\partial x}\right) \tag{191}$$

where $\mathbf{A}^\dagger$ is the *adjugate matrix* of $\mathbf{A}$. The (ij)-th element $A^\dagger_{ij}$ of the adjugate matrix $\mathbf{A}^\dagger$ is given by

$$A^\dagger_{ij} = (-1)^{i+j}\det\left(\mathbf{A}^{(ji)}\right) \tag{192}$$

(beware that the superscript (ji) of $\mathbf{A}^{(ji)}$ is *not* (ij) but it is "transposed") where $\mathbf{A}^{(ij)}$ (the superscript (ij) is *not* "transposed" here) denotes a matrix obtained by removing the i-th row and the j-th column of $\mathbf{A}$.

From the well-known identity

$$\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}^\dagger\mathbf{A} = \det(\mathbf{A})\mathbf{I} \tag{193}$$

(consult a linear algebra textbook for a proof), we can write the inverse matrix $\mathbf{A}^{-1}$ in terms of the adjugate matrix $\mathbf{A}^\dagger$ so that

$$\mathbf{A}^{-1} = \frac{\mathbf{A}^\dagger}{\det(\mathbf{A})} \tag{194}$$

if $\mathbf{A}$ is nonsingular so that $\det(\mathbf{A}) \neq 0$. Note also that the above identity (193) implies

$$\det(\mathbf{A}) = \sum_k A_{ik}A^\dagger_{ki} = \sum_k A^\dagger_{jk}A_{kj} \tag{195}$$

for any i and j. Substituting this identity (195) into the left hand side of (191) and noting that, from the definition (192) of the adjugate matrix, $A^\dagger_{ji}$ is independent of A_{ik} nor A_{kj} for any k, we have

$$\frac{\partial}{\partial x}\det(\mathbf{A}) = \sum_{ij}\left\{\frac{\partial}{\partial A_{ij}}\sum_k A_{ik}A^\dagger_{ki}\right\}\frac{\partial A_{ij}}{\partial x} = \sum_{ij}\left\{\frac{\partial}{\partial A_{ij}}\sum_k A^\dagger_{jk}A_{kj}\right\}\frac{\partial A_{ij}}{\partial x} \tag{196}$$

$$= \sum_{ij} A^\dagger_{ji}\frac{\partial A_{ij}}{\partial x} \tag{197}$$

$$= \mathrm{Tr}\left(\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^\dagger\right) = \mathrm{Tr}\left(\mathbf{A}^\dagger\frac{\partial \mathbf{A}}{\partial x}\right) \tag{198}$$

which proves the identity (191).

---

in terms of the eigenvalue decomposition so as instead to use the *singular value decomposition* or SVD (220) in order to show (C.22) for any nonsingular matrix $\mathbf{A}$ such that $\det(\mathbf{A}) \neq 0$. However, I would like to present yet another proof in terms of *Jacobi's formula* (191) here because it is more direct and general than the one in terms of the SVD.

I leave the proof of (C.22) in terms of the SVD as an exercise for the reader. Hint: The right hand side of (C.22) can be written as $\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\partial\mathbf{\Sigma}/\partial x\right)$ where $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ is the SVD of $\mathbf{A}$ because it follows from the *orthonormality* (C.37) of $\mathbf{U}$ that $\mathrm{Tr}(\mathbf{U}^{\mathrm{T}}\partial\mathbf{U}/\partial x) = 0$ and similarly for $\mathbf{V}$. Finally, note that the absolute determinant of $\mathbf{A}$ is equal to that of $\mathbf{\Sigma}$ and thus to the product of the singular values $\sigma_i > 0$ such that $\mathbf{\Sigma} = \mathrm{diag}\,(\sigma_i)$, i.e., $|\mathbf{A}| = |\mathbf{\Sigma}| = \prod_i \sigma_i$ where we have used (15).

Derivative of log absolute determinant Assuming that $\mathbf{A}$ is nonsingular so that $\det(\mathbf{A}) \neq 0$, we can evaluate the left hand side of (C.22) as

$$\frac{\partial}{\partial x} \ln |\mathbf{A}| = \frac{1}{\det(\mathbf{A})} \frac{\partial}{\partial x} \det(\mathbf{A}) \tag{199}$$

where we have used the notation (15) for $|\mathbf{A}|$. Substituting (191), we obtain

$$\frac{\partial}{\partial x} \ln |\mathbf{A}| = \mathrm{Tr}\left(\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}\right) = \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right) \tag{200}$$

where we have used (194).

Page 698

Equations (C.24) through (C.28): Since these derivatives are gradients of a scalar with respect to a matrix $\mathbf{A}$, the operator $\frac{\partial}{\partial \mathbf{A}}$ should read $\nabla_{\mathbf{A}}$ if we adopt the notation (190).

Matrix derivatives identities For example, (C.24) should read

$$\nabla_{\mathbf{A}} \mathrm{Tr}(\mathbf{AB}) = \nabla_{\mathbf{A}} \mathrm{Tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\right) = \mathbf{B}^{\mathrm{T}} \tag{201}$$

where we have used the transpose and the cyclic identities of the trace operator $\mathrm{Tr}(\cdot)$, i.e.,

$$\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}\left(\mathbf{A}^{\mathrm{T}}\right), \qquad \mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA}) \tag{202}$$

respectively. As described in the text, the identity (201) directly follows from (C.23). At this moment, I would like to point out an observation helpful for remembering (201). First, note that the gradient of a scalar with respect to a matrix $\mathbf{A}$ is, by definition (190), a matrix of the same dimensionality as $\mathbf{A}$. On the other hand, in order for the trace $\mathrm{Tr}(\mathbf{AB})$ to be meaningful, $\mathbf{B}$ must be of the same dimensionality as $\mathbf{A}^{\mathrm{T}}$. Thus, (201) passes the "dimensionality test," meaning that all the matrix operations in (201) are meaningful. Note also that (C.25) and (C.26) are special cases of (201).

Similarly, the gradient of the log (absolute) determinant (C.28) should read

$$\nabla_{\mathbf{A}} \ln |\mathbf{A}| = \mathbf{A}^{-\mathrm{T}} \tag{203}$$

where we have used (C.4) and defined

$$\mathbf{A}^{-\mathrm{T}} \equiv \left(\mathbf{A}^{\mathrm{T}}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}}. \tag{204}$$

Again, if we adopt the notation (15) for $|\mathbf{A}|$, we see that (203) holds for any nonsingular matrix $\mathbf{A}$ such that $\det(\mathbf{A}) \neq 0$.

The identity (203) can be shown by identifying x with A_{ij} in (200) where A_{ij} is the (ij)-th element of $\mathbf{A}$; and then making use of (C.23), which can be stated more suitably for our purpose here as

$$\mathrm{Tr}\left(\frac{\partial \mathbf{A}}{\partial A_{ij}}\mathbf{B}\right) = \mathrm{Tr}\left(\mathbf{B}\frac{\partial \mathbf{A}}{\partial A_{ij}}\right) = B_{ji}. \tag{205}$$

Here, the Jacobian matrix $\partial \mathbf{A}/\partial A_{ij}$ is, by definition (189), of the same dimensionality as $\mathbf{A}$ and such that only the (ij)-th element is one whereas all the other elements vanish so that

$$\frac{\partial \mathbf{A}}{\partial A_{ij}} = i \begin{pmatrix} & & j & \\ & \vdots & & \\ \cdots & 1 & \cdots & \\ & \vdots & & \end{pmatrix} \tag{206}$$

where the elements omitted or denoted by dots $(\cdots)$ are zero. Note that Svensén and Bishop (2009) effectively make use of (205) in the solution of Exercise 2.34.

In addition to the above mentioned matrix derivatives identities, I would suggest to include the following:

$$\nabla_{\mathbf{A}} \operatorname{Tr}\left(\mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{C}\right) = \mathbf{C}^{\mathrm{T}}\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{C}\mathbf{A}\mathbf{B} \tag{207}$$

$$\nabla_{\mathbf{A}} \operatorname{Tr}\left(\mathbf{A}^{-1}\mathbf{B}\right) = -\mathbf{A}^{-\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-\mathrm{T}}. \tag{208}$$

We use the identities (207) and (208), e.g., when we show (13.113) in Exercise 13.33 and (2.122) in Exercise 2.34, respectively. It should also be noted that (C.27) is a special case of (207).

The identity (207) can be shown as follows. Assuming that $\mathbf{B}$ and $\mathbf{C}$ are constants, we have

$$\frac{\partial}{\partial x}\left\{\mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{C}\right\} = \frac{\partial \mathbf{A}}{\partial x}\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{C} + \mathbf{A}\mathbf{B}\left(\frac{\partial \mathbf{A}}{\partial x}\right)^{\mathrm{T}}\mathbf{C}. \tag{209}$$

Taking the trace of the both sides gives

$$\frac{\partial}{\partial x}\operatorname{Tr}\left(\mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{C}\right) = \operatorname{Tr}\left(\frac{\partial \mathbf{A}}{\partial x}\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{C}\right) + \operatorname{Tr}\left(\frac{\partial \mathbf{A}}{\partial x}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\right) \tag{210}$$

where we have rearranged the factors inside the second trace $\operatorname{Tr}(\cdot)$ in the right hand side by making use of (202). We finally obtain (207) by identifying x with A_{ij} and making use of (205). The identity (208) follows similarly from (C.21).

Symmetric matrix derivatives So far, we have considered derivatives with respect to a matrix that is not necessary symmetric. However, matrices for which we take derivatives in order to, say, perform optimization (e.g., maximum likelihood) or evaluate expectations by making use of (23) are often symmetric. For example, the covariance $\mathbf{\Sigma}$ of the multivariate Gaussian distribution is symmetric positive definite. When we derive the maximum likelihood solution for $\mathbf{\Sigma}$ in Exercise 2.34, we ignore the symmetry constraint on $\mathbf{\Sigma}$ to calculate the derivatives of the log likelihood with respect to $\mathbf{\Sigma}$. The maximum likelihood solution $\mathbf{\Sigma}_{\mathrm{ML}}$ is obtained by solving necessary conditions that the derivatives should vanish, after which we find $\mathbf{\Sigma}_{\mathrm{ML}}$ to be symmetric positive definite. The fact that the solution $\mathbf{\Sigma}_{\mathrm{ML}}$ is symmetric is not a fortunate coincidence but a consequence of the symmetry in the necessary conditions solved. In fact, even if we had imposed the symmetry constraint on $\mathbf{\Sigma}$ in the first place, we would have obtained an equivalent set of equations to solve, giving the same solution. We can understand why this is the case by considering derivatives with respect to a symmetric matrix in more general terms as follows.

Let $\phi(\mathbf{A})$ be a scalar function of a square matrix $\mathbf{A}$ where $\mathbf{A} = (A_{ij})$ is not a symmetric matrix so that $A_{ij} \neq A_{ji}$. As usual, we write the gradient of $\phi(\mathbf{A})$ with respect to $\mathbf{A}$ as

$$\nabla_{\mathbf{A}}\phi(\mathbf{A}) = \left(\frac{\partial}{\partial A_{ij}}\phi(\mathbf{A})\right). \tag{211}$$

Suppose that we want to evaluate the gradient of $\phi(\mathbf{S})$ where $\mathbf{S} = (S_{ij})$ is a symmetric matrix so that $S_{ij} \equiv S_{ji}$. The derivative of $\phi(\mathbf{S})$ with respect to an off-diagonal element S_{ij} where $i \neq j$ consists of two derivatives through A_{ij} and A_{ji} so that

$$\frac{\partial}{\partial S_{ij}}\phi(\mathbf{S}) = \frac{\partial}{\partial A_{ij}}\phi(\mathbf{S}) + \frac{\partial}{\partial A_{ji}}\phi(\mathbf{S}) \tag{212}$$

where we have written

$$\frac{\partial}{\partial A_{ij}}\phi(\mathbf{S}) \equiv \left.\frac{\partial}{\partial A_{ij}}\phi(\mathbf{A})\right|_{\mathbf{A}=\mathbf{S}}. \tag{213}$$

The derivative of $\phi(\mathbf{S})$ with respect to an diagonal element S_{ii} is given by

$$\frac{\partial}{\partial S_{ii}}\phi(\mathbf{S}) = \frac{\partial}{\partial A_{ii}}\phi(\mathbf{S}). \tag{214}$$

Thus, we can write

$$\nabla_{\mathbf{S}}\phi(\mathbf{S}) = \nabla_{\mathbf{A}}\phi(\mathbf{S}) + \nabla_{\mathbf{A}}\phi(\mathbf{S})^{\mathrm{T}} - \mathrm{diag}\left(\nabla_{\mathbf{A}}\phi(\mathbf{S})\right) \tag{215}$$

where we have written

$$\nabla_{\mathbf{A}}\phi(\mathbf{S}) \equiv \left.\nabla_{\mathbf{A}}\phi(\mathbf{A})\right|_{\mathbf{A}=\mathbf{S}}. \tag{216}$$

For example, if $\mathbf{A}$ and $\mathbf{B}$ are both symmetric in (201), we have

$$\nabla_{\mathbf{A}}\,\mathrm{Tr}\left(\mathbf{AB}\right) = 2\mathbf{B} - \mathrm{diag}(\mathbf{B}). \tag{217}$$

The identity (215) is, however, not very useful in practice. A more useful observation can be made by considering equations obtained by setting the derivatives equal to zero. Specifically, it readily follows from (212) and (214) that, if $\nabla_{\mathbf{A}}\phi(\mathbf{S})$ is symmetric (which does hold, say, for the necessary conditions for $\mathbf{\Sigma}_{\mathrm{ML}}$ we mentioned above), we have

$$\nabla_{\mathbf{S}}\phi(\mathbf{S}) = \mathbf{O} \quad \Longleftrightarrow \quad \nabla_{\mathbf{A}}\phi(\mathbf{S}) = \mathbf{O} \tag{218}$$

which implies that we can solve $\nabla_{\mathbf{S}}\phi(\mathbf{S}) = \mathbf{O}$ without the symmetry constraint on $\mathbf{S}$, i.e., by simply solving $\nabla_{\mathbf{A}}\phi(\mathbf{S}) = \mathbf{O}$ and then obtain a solution $\mathbf{S}$ that is indeed symmetric.

When we evaluate expectations by making use of (23), we consider equations obtained by setting the expected derivatives equal to zero. With much the same discussion as above, if $\mathbb{E}\left[\nabla_{\mathbf{A}}\phi(\mathbf{S})\right]$ is symmetric, we have

$$\mathbb{E}\left[\nabla_{\mathbf{S}}\phi(\mathbf{S})\right] = \mathbf{O} \quad \Longleftrightarrow \quad \mathbb{E}\left[\nabla_{\mathbf{A}}\phi(\mathbf{S})\right] = \mathbf{O}. \tag{219}$$

It should be noted that, since the score function (22) occurring in (23) is defined only for independent parameters, if the parameters of interest are, say, a symmetric matrix (e.g., the scale matrix $\mathbf{W}$ of the Wishart distribution is symmetric positive definite), we must, strictly speaking, impose the symmetry constraint on the parameters. The equivalence relation (219) allows us, if $\mathbb{E}\left[\nabla_{\mathbf{A}}\phi(\mathbf{S})\right]$ is symmetric, to safely ignore the symmetry constraint on $\mathbf{S}$ and use $\mathbb{E}\left[\nabla_{\mathbf{A}}\phi(\mathbf{S})\right] = \mathbf{O}$ instead.

Page 700

Paragraph 2, Line –1: The determinant of the orthogonal matrix $\mathbf{U}$ can be either positive or negative so that we should write $\det(\mathbf{U}) = \pm 1$ (which is, if the notation (15) is adopted, equivalent to $|\mathbf{U}| = 1$). Although it is possible to take $\mathbf{U}$ such that $\det(\mathbf{U}) = 1$ (one can flip the sign of $\det(\mathbf{U})$ by, say, flipping the sign of any one of the eigenvectors $\{\mathbf{u}_i\}$), there is no point in doing so in practice theoretically nor numerically. In fact, it is easy to see that the following discussion remains valid provided that $\mathbf{U}$ is orthogonal so that we have (C.37) but not necessarily that $\det(\mathbf{U}) = 1$. Moreover, most software implementations of symmetric eigenvalue decomposition only guarantee that $\mathbf{U}$ is orthogonal so that $\det(\mathbf{U}) = \pm 1$.

Singular value decomposition In the special case where the matrix $\mathbf{A}$ is symmetric positive semidefinite or $\mathbf{A} \succeq 0$, we can identify the eigenvalue decomposition (C.43) with the *singular value decomposition* or SVD (Press et al., 1992; Golub and Van Loan, 2013) so that we can use an SVD routine to compute the eigenvalue decomposition of $\mathbf{A}$. The SVD is generally defined for any real matrix $\mathbf{P}$ not necessarily square, say, of dimensionality $M \times N$, so that the SVD of $\mathbf{P}$ is given by

$$\mathbf{P} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}} = \sum_{i=1}^{R} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \tag{220}$$

where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_M)$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$ are orthogonal matrices of dimensionalities $M \times M$ and $N \times N$, respectively; $\boldsymbol{\Sigma}$ is an $M \times N$ diagonal matrix with nonnegative diagonal elements, called the *singular values*, $\sigma_1 \geqslant \cdots \geqslant \sigma_R \geqslant 0$ arranged in descending order; and $R \leqslant \min(M, N)$ is the rank of $\mathbf{P}$. Note again that $\mathbf{U}$ and $\mathbf{V}$ are only guaranteed to be orthogonal so that $\det(\mathbf{U}) = \pm 1$ and $\det(\mathbf{V}) = \pm 1$.

Page 700

The text following (C.41): The multiplication by $\mathbf{U}$ can be interpreted as a rotation, a reflection, or a combination of the two.

Page 705

Equation (D.8): It would be helpful if we make it clear that the left hand side of (D.8) corresponds to the functional derivative so that we should modify (D.8) as

$$\frac{\delta F}{\delta y(x)} \equiv \frac{\partial G}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial G}{\partial y'}\right) = 0. \tag{221}$$

Page 705

Paragraph –1, Line 1: Despite the statement, it is not that straightforward to extend the results obtained here to higher dimensions. Although such an extension is not required in PRML, it is useful when we analyze a particular type of constrained optimization problem commonly found in computer vision applications such as *optical flow* (Horn and Schunck, 1981). Here, I would like to consider an extension of the calculus of variations to a system of D-dimensional Cartesian coordinates $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}} \in \mathbb{R}^D$ and find the form of the functional derivative as well as a more general boundary condition for such a derivative to be

well-defined. To this end, we first review some identities concerning the *divergence* (Feynman et al., 1964). The divergence of a vector field

$$\mathbf{p}(\mathbf{x}) = \begin{pmatrix} p_1(\mathbf{x}) \\ \vdots \\ p_D(\mathbf{x}) \end{pmatrix} \in \mathbb{R}^D \tag{222}$$

is a scalar field of the form

$$\operatorname{div} \mathbf{p} = \sum_{i=1}^{D} \frac{\partial p_i}{\partial x_i} \equiv \nabla \cdot \mathbf{p} \tag{223}$$

where we have omitted the coordinates $\mathbf{x}$ in the function arguments to keep the notation uncluttered. For a differentiable vector field $\mathbf{p}(\mathbf{x})$ defined on some volume $\Omega \subset \mathbb{R}^D$, the *divergence theorem* (Feynman et al., 1964) states that

$$\int_\Omega \operatorname{div} \mathbf{p} \, \mathrm{d}V = \oint_{\partial\Omega} \mathbf{p} \cdot \mathbf{n} \, \mathrm{d}S \tag{224}$$

where the left hand side is the volume integral over the volume Ω; the right hand side is the surface integral over its boundary $\partial\Omega$; and $\mathbf{n}(\mathbf{x})$ is the outward unit normal vector of $\partial\Omega$. Assuming that the coordinates $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$ are Cartesian, we can write the volume element as $\mathrm{d}V = \mathrm{d}x_1 \cdots \mathrm{d}x_D \equiv \mathrm{d}\mathbf{x}$ and the inner product as $\mathbf{p} \cdot \mathbf{n} = \mathbf{p}^{\mathrm{T}}\mathbf{n}$. Making use of the divergence theorem (224) together with the following identity

$$\operatorname{div}(\phi\mathbf{p}) = \nabla\phi^{\mathrm{T}}\mathbf{p} + \phi \operatorname{div} \mathbf{p} \tag{225}$$

we obtain a multidimensional version of the "integration by parts" formula

$$\int_\Omega \nabla\phi^{\mathrm{T}}\mathbf{p} \, \mathrm{d}\mathbf{x} = \oint_{\partial\Omega} \phi\mathbf{p}^{\mathrm{T}}\mathbf{n} \, \mathrm{d}S - \int_\Omega \phi \operatorname{div} \mathbf{p} \, \mathrm{d}\mathbf{x}. \tag{226}$$

Let us now consider a functional of the form

$$E\left[u(\mathbf{x})\right] = \int_\Omega L\left(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x})\right) \mathrm{d}\mathbf{x} \tag{227}$$

where $u(\mathbf{x}) \in \mathbb{R}$ is a function (scalar field) defined over some volume $\Omega \subset \mathbb{R}^D$ and $L(\mathbf{x}, f, \mathbf{g}) \in \mathbb{R}$ is a function of $\mathbf{x} \in \Omega$, $f \in \mathbb{R}$, and $\mathbf{g} \in \mathbb{R}^D$. Thus, the functional $E\left[u(\mathbf{x})\right] \in \mathbb{R}$ maps $u(\mathbf{x})$ to a real number. As in the ordinary calculus, we can define the derivative of a functional according to the *calculus of variations* (Feynman et al., 1964; Bishop, 2006). In order to find the form of the functional derivative, we consider how $E\left[u(\mathbf{x})\right]$ varies upon a small change $\epsilon\eta(\mathbf{x})$ in $u(\mathbf{x})$ where $\eta(\mathbf{x})$ is the "direction" of the change and ϵ is some small constant. The first-order variation of $E\left[u(\mathbf{x})\right]$ in the direction of $\eta(\mathbf{x})$ can be evaluated as

$$\begin{aligned}
\delta E[u;\eta] &\equiv \lim_{\epsilon\to 0} \frac{1}{\epsilon} \left\{E[u+\epsilon\eta] - E[u]\right\} && (228)\\
&= \lim_{\epsilon\to 0} \frac{1}{\epsilon} \int_\Omega \left\{L\left(\mathbf{x}, u+\epsilon\eta, \nabla(u+\epsilon\eta)\right) - L\left(\mathbf{x}, u, \nabla u\right)\right\} \mathrm{d}\mathbf{x} && (229)\\
&= \int_\Omega \left\{\eta \frac{\partial L}{\partial f} + \nabla\eta^{\mathrm{T}} \nabla_{\mathbf{g}} L\right\} \mathrm{d}\mathbf{x} && (230)
\end{aligned}$$

where we have assumed that $L(\mathbf{x}, f, \mathbf{g})$ is differentiable with respect to both f and $\mathbf{g}$; and we have written

$$\frac{\partial L}{\partial f} \equiv \frac{\partial}{\partial f} L\left(\mathbf{x}, u, \nabla u\right), \qquad \nabla_{\mathbf{g}} L \equiv \nabla_{\mathbf{g}} L\left(\mathbf{x}, u, \nabla u\right). \tag{231}$$

By making use of the multidimensional integration by parts (226), we can integrate the second term in the right hand side of (230), giving

$$\delta E[u;\eta] = \int_\Omega \eta \left\{ \frac{\partial L}{\partial f} - \operatorname{div}\left(\nabla_{\mathbf{g}} L\right) \right\} \mathrm{d}\mathbf{x} + \oint_{\partial\Omega} \eta \nabla_{\mathbf{g}} L^{\mathrm{T}} \mathbf{n} \, \mathrm{d}S. \tag{232}$$

In order for the functional derivative to be well-defined, we assume the surface integral term in the variation (232) to vanish so that we have the following boundary condition

$$\oint_{\partial\Omega} \eta \nabla_{\mathbf{g}} L^{\mathrm{T}} \mathbf{n} \, \mathrm{d}S = 0. \tag{233}$$

The boundary condition (233) holds if

$$\eta(\mathbf{x}) = 0 \tag{234}$$

or

$$\nabla_{\mathbf{g}} L^{\mathrm{T}} \mathbf{n}(\mathbf{x}) = 0 \tag{235}$$

for all $\mathbf{x} \in \partial\Omega$. The first condition (234) holds if we assume the *Dirichlet boundary condition* for $u(\mathbf{x})$

$$u(\mathbf{x}) = u_0(\mathbf{x}) \tag{236}$$

where $\mathbf{x} \in \partial\Omega$, i.e., $u(\mathbf{x})$ is assumed to be fixed to some value $u_0(\mathbf{x})$ at the boundary $\partial\Omega$ and so is $u(\mathbf{x}) + \epsilon\eta(\mathbf{x})$ in (228), implying (234). Another common boundary condition for $u(\mathbf{x})$ is the *Neumann boundary condition*

$$\nabla u(\mathbf{x})^{\mathrm{T}} \mathbf{n}(\mathbf{x}) = 0 \tag{237}$$

where $\mathbf{x} \in \partial\Omega$. The Neumann boundary condition (237) is implied by the second condition (235) for the optical-flow energy functional as we shall see shortly. Having assumed that the boundary condition (233) holds, we can write the first order variation (232) in the form

$$\delta E[u;\eta] = \int_\Omega \eta \frac{\partial E}{\partial u(\mathbf{x})} \mathrm{d}\mathbf{x} \tag{238}$$

where we have written

$$\frac{\partial E}{\partial u(\mathbf{x})} \equiv \frac{\partial L}{\partial f} - \operatorname{div}\left(\nabla_{\mathbf{g}} L\right). \tag{239}$$

The volume integral in the right hand side of (238) can be seen as the inner product between $\eta(\mathbf{x})$ and $\partial E/\partial u(\mathbf{x})$, from which we conclude that the quantity $\partial E/\partial u(\mathbf{x})$ is what should be called the functional derivative.[12] A stationary point of a functional $E[u(\mathbf{x})]$ is a function $u(\mathbf{x})$ such that the variation $\delta E[u;\eta]$ vanishes in any direction $\eta(\mathbf{x})$ and thus satisfies the *Euler-Lagrange equation* given by

$$\frac{\partial E}{\partial u(\mathbf{x})} = 0. \tag{240}$$

Finally, we present an application of the multidimensional calculus of variations to a dense motion analysis technique called optical flow in the following. Suppose that, given a pair

[12] Here we use a notation for the functional derivative that is different from the one used in PRML. The notation $\partial E/\partial u(\mathbf{x})$ employed here is more like an ordinary derivative and can be extended to the case of a vector field $\mathbf{u}(\mathbf{x})$ analogously to the gradient as we shall see in (245).

of (grayscale) images $I_0(\mathbf{x})$ and $I_1(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^2$ that are taken at some discrete time steps $t = 0$ and $t = 1$, respectively, we wish to find a motion vector field from $I_0(\mathbf{x})$ to $I_1(\mathbf{x})$

$$\mathbf{u}(\mathbf{x}) = \begin{pmatrix} u(\mathbf{x}) \\ v(\mathbf{x}) \end{pmatrix} \tag{241}$$

defined over $\mathbf{x} \in \Omega \subset \mathbb{R}^2$. Horn and Schunck (1981) sought for $\mathbf{u}(\mathbf{x})$ that minimizes an energy functional that takes essentially the same form as

$$J[\mathbf{u}(\mathbf{x})] = J_{\text{data}}[\mathbf{u}(\mathbf{x})] + \alpha J_{\text{smooth}}[\mathbf{u}(\mathbf{x})] \tag{242}$$

where

$$J_{\text{data}}[\mathbf{u}(\mathbf{x})] = \frac{1}{2} \int_\Omega \left(I_1(\mathbf{x} + \mathbf{u}(\mathbf{x})) - I_0(\mathbf{x}) \right)^2 \mathrm{d}\mathbf{x} \tag{243}$$

$$J_{\text{smooth}}[\mathbf{u}(\mathbf{x})] = \frac{1}{2} \int_\Omega \left(\|\nabla u(\mathbf{x})\|^2 + \|\nabla v(\mathbf{x})\|^2 \right) \mathrm{d}\mathbf{x}. \tag{244}$$

Here, the domain Ω is assumed to be continuous and is typically rectangular. We call the first term $J_{\text{data}}[\mathbf{u}(\mathbf{x})]$ in (242) the data-fidelity term; the second term $J_{\text{smooth}}[\mathbf{u}(\mathbf{x})]$ the smoothness (regularization) term; and the coefficient α the regularization parameter. According to the multidimensional calculus of variations, a stationary point of the optical-flow energy functional (242) satisfies Euler-Lagrange equations of the form

$$\nabla_{\mathbf{u}(\mathbf{x})} J \equiv \begin{pmatrix} \partial J / \partial u(\mathbf{x}) \\ \partial J / \partial v(\mathbf{x}) \end{pmatrix} = \nabla_{\mathbf{u}} \left\{ \frac{\varepsilon\left(\mathbf{x}, \mathbf{u}(\mathbf{x})\right)^2}{2} \right\} - \alpha \begin{pmatrix} \operatorname{div}\left(\nabla u(\mathbf{x})\right) \\ \operatorname{div}\left(\nabla v(\mathbf{x})\right) \end{pmatrix} = \mathbf{0} \tag{245}$$

where we have written

$$\varepsilon\left(\mathbf{x}, \mathbf{u}\right) = I_1(\mathbf{x} + \mathbf{u}) - I_0(\mathbf{x}). \tag{246}$$

For the functional derivatives $\partial J / \partial u(\mathbf{x})$ and $\partial J / \partial v(\mathbf{x})$ to be well-defined, let us assume the boundary condition given by (235) for each functional derivative, which implies the Neumann boundary condition for $\mathbf{u}(\mathbf{x})$, i.e.,

$$\nabla u(\mathbf{x})^{\mathrm{T}} \mathbf{n}(\mathbf{x}) = 0, \qquad \nabla v(\mathbf{x})^{\mathrm{T}} \mathbf{n}(\mathbf{x}) = 0 \tag{247}$$

for all $\mathbf{x} \in \partial\Omega$ where $\partial\Omega$ is the boundary of Ω and $\mathbf{n}(\mathbf{x})$ is the outward unit normal vector of $\partial\Omega$. Thus, solving the above Euler-Lagrange equations (245) with the Neumann boundary condition (247), we obtain the desired motion vector field $\mathbf{u}(\mathbf{x})$. The Euler-Lagrange equations given by (245) are *elliptic partial differential equations* (elliptic PDEs) and can be solved numerically by a type of relaxation method such as the Gauss-Seidel method or the (weighted) Jacobi method or by a more efficient *multigrid* technique (Press et al., 1992; Briggs et al., 2000).

Page 708

Equation (E.3): The right hand side should be a zero vector $\mathbf{0}$ instead of a scalar zero 0.

Page 708

The text after (E.4): $\nabla_{\mathbf{x}} L = 0$ should read $\nabla_{\mathbf{x}} L = \mathbf{0}$ (the right hand side should be a zero vector $\mathbf{0}$).

Page 709

Paragraph –2, Line 5: $\nabla f(\mathbf{x}) = 0$ should read $\nabla f(\mathbf{x}) = \mathbf{0}$ (the right hand side should be a zero vector $\mathbf{0}$).

Page 716

Column 1, Entry –1: "The Feynman Lectures of Physics" should read "The Feynman Lectures on Physics."

Page 717

Column 2, Entry 7: "John Hopkins University Press" should read "The Johns Hopkins University Press."

References

Abramowitz, M. and I. A. Stegun (Eds.) (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, U.S. Department of Commerce. 14

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (Third ed.). Wiley. 34, 35

Anderson, T. W. and I. Olkin (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra and Its Applications 70*, 147–171. 7

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 1, 45

Briggs, W. L., V. E. Henson, and S. F. McCormick (2000). *A Multigrid Tutorial* (Second ed.). SIAM. 47

Feynman, R. P., R. B. Leighton, and M. Sands (1964). *The Feynman Lectures on Physics*, Volume 2. Addison-Wesley. 45

Golub, G. H. and C. F. Van Loan (2013). *Matrix Computations* (Fourth ed.). The Johns Hopkins University Press. 6, 44

Horn, B. K. P. and B. G. Schunck (1981). Determining optical flow. *Artificial Intelligence 17*(1), 185–203. 44, 47

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86. 4

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. 3, 4

Magnus, J. R. and H. Neudecker (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (Third ed.). Wiley. http://www.janmagnus.nl/misc/mdc2007-3rdedition. 7

Minka, T. P. (2000). Old and new matrix algebra useful for statistics. https://tminka.github.io/papers/matrix/. 36, 37

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press. 32

Olver, F. W. J., A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, and B. V. Saunders (Eds.) (2016). *NIST Digital Library of Mathematical Functions*. National Institute of Standards and Technology, U.S. Department of Commerce. http://dlmf.nist.gov/ (Release 1.0.14 of 2016-12-21). 2, 14, 34

Press, W. M., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing* (Second ed.). Cambridge University Press. 14, 17, 44, 47

Svensén, M. and C. M. Bishop (2005). Robust Bayesian mixture modelling. *Neurocomputing 64*, 235–252. 35

Svensén, M. and C. M. Bishop (2009). Pattern Recognition and Machine Learning: Solutions to the exercises (web edition). https://www.microsoft.com/en-us/research/people/cmbishop/#prml-book. 24, 42

Svensén, M. and C. M. Bishop (2011). Pattern Recognition and Machine Learning: Errata and additional comments. https://www.microsoft.com/en-us/research/people/cmbishop/#prml-book. 1, 22, 23, 26, 29

Szeliski, R. (2010). *Computer Vision: Algorithms and Applications*. Springer. 14

Tipping, M. E. and A. C. Faul (2003). Fast marginal likelihood maximisation for sparse Bayesian models. In C. M. Bishop and B. J. Frey (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, Florida. 20