PRML Errata

Yousuke Takada

October 23, 2015

Preface

This report communicates some possible errata for PRML (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, I follow the notational conventions adopted by Svensén and Bishop (2011). I have also included in this report some suggestions for improving the readability.

Corrections

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 $\mathrm{H}[p]$.

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, Equation (2.52) and the following text should read

$$\mathbf{y} = \mathbf{U}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) \tag{1}$$

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 (1) instead of (2.52), the Jacobian matrix $\mathbf{J} = (J_{ij})$ we require is simply given by $\mathbf{U}$. Equation (2.53) should read

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ij} \tag{2}$$

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

$$|\mathbf{J}|^2 = |\mathbf{U}|^2 = \left|\mathbf{U}^{\mathrm{T}}\right| |\mathbf{U}| = \left|\mathbf{U}^{\mathrm{T}}\mathbf{U}\right| = |\mathbf{I}| = 1. \tag{3}$$

[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.

Page 81

Line −1: Since the determinant of the Jacobian, which is an orthonormal matrix here, can be negative, 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; the factor $|\mathbf{J}|$ should read $|\det(\mathbf{J})|$. Note that we cannot write $||\mathbf{J}||$ to mean $|\det(\mathbf{J})|$ because it 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 has been 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. An alternative solution to this problem would be to explicitly define

$$|\mathbf{A}| \equiv |\det(\mathbf{A})| \tag{4}$$

for any square matrix $\mathbf{A}$. This notation is mostly consistent because we have $|\mathbf{A}| = \det(\mathbf{A})$ for a positive semidefinite matrix $\mathbf{A}$ and most other matrices for which we take determinants are positive (semi)definite in PRML.

Page 102

Equation (2.155): Although an interpretation for the parameters of the gamma distribution (2.146) has been made, one for the parameters of the Wishart distribution (2.155) is not given here nor in Exercise 2.45. In order to give one, 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 $\mathbf{\Sigma}$ or, equivalently, the precision matrix $\mathbf{\Lambda} \equiv \mathbf{\Sigma}^{-1}$. The likelihood $p(\mathbf{X}|\mathbf{\Lambda})$ in terms of the precision $\mathbf{\Lambda}$ is given by

$$p(\mathbf{X}|\mathbf{\Lambda}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{\Lambda}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n \middle| \mathbf{0}, \mathbf{\Lambda}^{-1}\right). \tag{5}$$

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

$$p(\mathbf{\Lambda}) = \mathcal{W}(\mathbf{\Lambda}|\mathbf{W}_0, \nu_0) \tag{6}$$

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

$$\begin{aligned} p(\mathbf{\Lambda}|\mathbf{X}) &\propto p(\mathbf{X}|\mathbf{\Lambda})\, p(\mathbf{\Lambda}) && (7)\\ &\propto |\mathbf{\Lambda}|^{N/2} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N} \mathbf{x}_n^{\mathrm{T}} \mathbf{\Lambda} \mathbf{x}_n\right\} |\mathbf{\Lambda}|^{(\nu_0 - D - 1)/2} \exp\left\{-\frac{1}{2}\operatorname{Tr}\left(\mathbf{W}_0^{-1}\mathbf{\Lambda}\right)\right\} && (8)\\ &= |\mathbf{\Lambda}|^{(\nu_N - D - 1)/2} \exp\left\{-\frac{1}{2}\operatorname{Tr}\left(\mathbf{W}_N^{-1}\mathbf{\Lambda}\right)\right\} && (9) \end{aligned}$$

where

$$\nu_N = \nu_0 + N \tag{10}$$

$$\mathbf{W}_N^{-1} = \mathbf{W}_0^{-1} + \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}. \tag{11}$$

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

$$p\left(\mathbf{\Lambda}|\mathbf{X}\right) = \mathcal{W}\left(\mathbf{\Lambda}|\mathbf{W}_N, \nu_N\right). \tag{12}$$

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 (6) as the number of "effective" prior observations. The N observations also contribute $N\mathbf{\Sigma}_{\mathrm{ML}}$ to the inverse of the scale matrix $\mathbf{W}$ where $\mathbf{\Sigma}_{\mathrm{ML}}$ is the maximum likelihood estimate for the covariance of the observations given by

$$\mathbf{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}. \tag{13}$$

This suggests an interpretation of $\mathbf{W}$ in terms of $\mathbf{\Sigma} \equiv (\nu\mathbf{W})^{-1}$. More specifically, we can interpret $\mathbf{\Sigma}_0 = (\nu_0\mathbf{W}_0)^{-1}$ as the covariance of the ν_0 "effective" prior observations. Note that this interpretation is in accordance with another observation that the expectation of $\mathbf{\Lambda}$ taken over the prior (6) is indeed given by $\mathbb{E}\left[\mathbf{\Lambda}\right] = \nu_0\mathbf{W}_0 = \mathbf{\Sigma}_0^{-1}$ where we have used (B.80).

Page 102

Line −2: 'Gamma' should read 'gamma' (without capitalization).

Page 104

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

Page 141

Equation (3.13): The use of the gradient operator is not consistent here. As in Equation (2.224), the gradient (of a scalar function) is a column vector so that Equation (3.13) should read[2]

$$\nabla \ln p\left(\mathbf{t}|\mathbf{w}, \beta\right) = \beta\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\left(\mathbf{x}_n\right)\right\}\boldsymbol{\phi}\left(\mathbf{x}_n\right). \tag{14}$$

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 will 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. Thus, Equation (3.14) should read

$$\mathbf{0} = \sum_{n=1}^{N} t_n\boldsymbol{\phi}\left(\mathbf{x}_n\right) - \left(\sum_{n=1}^{N}\boldsymbol{\phi}\left(\mathbf{x}_n\right)\boldsymbol{\phi}\left(\mathbf{x}_n\right)^{\mathrm{T}}\right)\mathbf{w} \tag{15}$$

where we have used the gradient of the form (14) instead of (3.13).

[2] I have not got the right typeface for the data vector $\left(t_1, \ldots, t_N\right)^{\mathrm{T}}$. See "Mathematical notation" of PRML.

Page 146

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

Page 166

The second paragraph, 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, Equation (3.117) should read

$$\frac{\partial}{\partial \alpha} \ln |\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1} \frac{\partial}{\partial \alpha} \mathbf{A}\right). \tag{16}$$

Page 207

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

$$\mathbf{w}^{\mathrm{new}} = \mathbf{w}^{\mathrm{old}} - \left[\mathbf{H}\left(\mathbf{w}^{\mathrm{old}}\right)\right]^{-1} \nabla E\left(\mathbf{w}^{\mathrm{old}}\right) \tag{17}$$

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

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N} y_{nj} (I_{kj} - y_{nk}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}. \tag{18}$$

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

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_K \end{pmatrix} \tag{19}$$

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

$$\nabla_{\mathbf{w}} E = \begin{pmatrix} \nabla_{\mathbf{w}_1} E \\ \vdots \\ \nabla_{\mathbf{w}_K} E \end{pmatrix}, \qquad \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} E = \begin{pmatrix} \nabla_{\mathbf{w}_1} \nabla_{\mathbf{w}_1} E & \cdots & \nabla_{\mathbf{w}_1} \nabla_{\mathbf{w}_K} E \\ \vdots & \ddots & \vdots \\ \nabla_{\mathbf{w}_K} \nabla_{\mathbf{w}_1} E & \cdots & \nabla_{\mathbf{w}_K} \nabla_{\mathbf{w}_K} E \end{pmatrix} \tag{20}$$

respectively.

Page 239

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

Page 251

The second paragraph: The approximation of the form (5.84) is usually referred to as the *Gauss–Newton* approximation, but not *Levenberg–Marquardt*. The Levenberg–Marquardt method is a method that improves the numerical stability of (Gauss–)Newton iterations by correcting the Hessian matrix so as to be more diagonal dominant (Press et al., 1992).

Page 275

The text after Equation (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

$$s^2(\mathbf{x}) = \mathbb{E}\left[\mathrm{Tr}\left\{(\mathbf{t} - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right])\left(\mathbf{t} - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\right)^{\mathrm{T}}\right\}\middle|\mathbf{x}\right] \tag{21}$$

$$= \sum_{k=1}^{K} \pi_k(\mathbf{x})\,\mathrm{Tr}\left\{\sigma_k^2(\mathbf{x})\mathbf{I} + \left(\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\right)\left(\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\right)^{\mathrm{T}}\right\} \tag{22}$$

$$= \sum_{k=1}^{K} \pi_k(\mathbf{x})\left\{L\sigma_k^2(\mathbf{x}) + \left\|\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\right\|^2\right\} \tag{23}$$

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

Page 295

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

Page 318

The text before Equation (6.93) as well as Equations (6.93) and (6.94): 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

$$\frac{\partial \mathbf{a}_N^\star}{\partial \theta_j} = \frac{\partial \mathbf{C}_N}{\partial \theta_j}\left(\mathbf{t}_N - \boldsymbol{\sigma}_N\right) - \mathbf{C}_N\mathbf{W}_N\frac{\partial \mathbf{a}_N^\star}{\partial \theta_j} \tag{24}$$

where the derivatives are Jacobians defined by (C.16) for a vector and analogously by (86) for a matrix. Rearranging (24) then gives

$$\frac{\partial \mathbf{a}_N^\star}{\partial \theta_j} = \left(\mathbf{I} + \mathbf{C}_N\mathbf{W}_N\right)^{-1}\frac{\partial \mathbf{C}_N}{\partial \theta_j}\left(\mathbf{t}_N - \boldsymbol{\sigma}_N\right). \tag{25}$$

Page 355

Equation (7.117): The typeface of the vector $\mathbf{y}$ in (7.117) should be that in (7.110).

Page 414

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 Equation (9.13): It should be noted for clarity that, as the prior $p(\mathbf{z})$ over $\mathbf{z}$ is a multinomial distribution (9.10), the posterior $p(\mathbf{z}|\mathbf{x})$ over $\mathbf{z}$ given $\mathbf{x}$ is again a multinomial of the form

$$p(\mathbf{z}|\mathbf{x}) = \prod_{k=1}^{K} \gamma_k^{z_k} \tag{26}$$

where we have written $\gamma_k \equiv \gamma(z_k)$, which can be directly confirmed by inspecting the functional form of the joint distribution

$$p(\mathbf{z})\, p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \left\{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}^{z_k}. \tag{27}$$

This observation helps the reader to understand that evaluating the responsibilities $\gamma(z_k)$ indeed corresponds to the E step of the general EM algorithm.

Page 434

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

$$\mathcal{N}\left(\mathbf{x}_n \middle| \mathbf{x}_n, \sigma_j^2 \mathbf{I}\right) = \frac{1}{\left(2\pi\sigma_j^2\right)^{D/2}}. \tag{28}$$

Page 435

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

Page 465

Equations (10.6) and (10.7): In PRML, Equation (10.6) will be later recognized as "a negative Kullback-Leibler divergence between $q_j(\mathbf{Z}_j)$ and $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$" (Page 465, Line -2). However, there is no point in taking a 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. We actually do not have to introduce the probability $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$ other than $q_j^\star(\mathbf{Z}_j)$. Specifically, we can rewrite Equations (10.6) and (10.7) into

$$\begin{aligned}
\mathcal{L}(q) &= \ldots && (29)\\
&= \int q_j \ln q_j^\star \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \mathrm{d}\mathbf{Z}_j + \text{const} && (30)\\
&= -\,\mathrm{KL}\left(q_j \middle\| q_j^\star\right) + \text{const} && (31)
\end{aligned}$$

where we have defined a new distribution $q_j^\star(\mathbf{Z}_j)$ by the relation

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right] + \text{const}. \tag{32}$$

It directly follows from (31) 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 Equation (10.8): The latent variable $\mathbf{z}_i$ should read $\mathbf{Z}_i$.

Page 465

Line -1: If we adopt the representation (31), the probability $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$ should read $q_j^\star(\mathbf{Z}_j)$.

Page 466

Line 1: Again, $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$ should read $q_j^\star(\mathbf{Z}_j)$. The sentence "Thus we obtain..." should read "Thus we see that we have already obtained a general expression for the optimal solution in (32)."

Page 468

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

Page 478

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

$$\nu_k = \nu_0 + N_k. \tag{33}$$

A quick check for the correctness of the re-estimation equations would be to consider a 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 (33).

Page 489

Equation (10.107): The expectations $\mathbb{E}_\alpha [\ln q(\mathbf{w})]_{\mathbf{w}}$ and $\mathbb{E}[\ln q(\alpha)]$ should read $\mathbb{E}_{\mathbf{w}}[\ln q(\mathbf{w})]$ and $\mathbb{E}_\alpha[\ln q(\alpha)]$, respectively, where the expectation $\mathbb{E}_{\mathbf{Z}}[\cdot]$ is taken over $q(\mathbf{Z})$.

Page 489

Equations (10.108) through (10.112): The expectations are notationally inconsistent with (1.36); they should be of the forms shown in (10.107) or the ones corrected as above.

Page 490

The third paragraph, Line 2: A comma (,) should be inserted after the ellipsis so that the range of n should read: $n = 1, \ldots, N$.

Page 496

Equation (10.140): In order to be consistent with the mathematical notations in PRML, the differential operator d in (10.140) should be 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 501

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

Page 503

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

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).[3] 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

$$p(\mathcal{D}) \simeq (2\pi v)^{D/2} \exp(B/2) \prod_{n=1}^{N} \left\{ s_n (2\pi v_n)^{-D/2} \right\} \tag{34}$$

where

$$B = \frac{\mathbf{m}^{\mathrm{T}}\mathbf{m}}{v} - \sum_{n=1}^{N} \frac{\mathbf{m}_n^{\mathrm{T}}\mathbf{m}_n}{v_n} \tag{35}$$

$$v^{-1} = \sum_{n=1}^{N} v_n^{-1} \tag{36}$$

$$v^{-1}\mathbf{m} = \sum_{n=1}^{N} v_n^{-1}\mathbf{m}_n. \tag{37}$$

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 $\hat{p}(\mathbf{x})$ 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

$$\hat{p}(\mathbf{x}) \propto q^{\backslash b}(\mathbf{x}) f_b(x_2, x_3) = \ldots \tag{38}$$

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

Page 515

The text after Equation (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 $\tilde{f}_j(\boldsymbol{\theta}_j)$.

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

Page 554

Equation (11.72), Line −2: If we write the expectation $\mathbb{E}_{\mathbf{z}}[\cdot]$ taken over some given distribution $q(\mathbf{z})$ explicitly as $\mathbb{E}_{q(\mathbf{z})}[\cdot]$, the expectation in the last line but one of (11.72) should read

$$\mathbb{E}_{p_G(\mathbf{z})}\left[\exp\left(-E(\mathbf{z})+G(\mathbf{z})\right)\right] \tag{39}$$

where we have written the argument $\mathbf{z}$ for $E(\mathbf{z})$ and $G(\mathbf{z})$ for clarity.

Page 556

Exercise 11.7: 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 Equation (12.12): The derivative we consider here is that with respect to b_j (*not* that with respect to b_i).

Page 564

The text after Equation (12.15): The zero should be a zero vector so that we have $\mathbf{u}_j = \mathbf{0}$.

Page 575

The third paragraph, 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 $\boldsymbol{\mu} = \bar{\mathbf{x}}$ into (12.53).

Page 578

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

Page 579

Line 5: The eigendecomposition requires $O(D^3)$ computation*s*.

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

The first paragraph, Line −5: The text "our predictions ..." should read: "our predictions for $\mathbf{x}_{n+1}$ depend on all the previous observations."

Page 620

The second paragraph and the following (unlabeled) equation: The last sentence before the equation and the equation each should terminate with a period (.).

Page 621

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

Page 622

The second paragraph, Line −1: "we see" should be omitted.

Page 623

The first paragraph, Line −2: z_{nk} should read $z_{n-1,k}$.

Page 631

Equation (13.73): The equation should read

$$\sum_{r=1}^{R} \ln \left\{ \frac{p\left(\mathbf{X}_r|\boldsymbol{\theta}_{m_r}\right) p(m_r)}{\sum_{l=1}^{M} p\left(\mathbf{X}_r|\boldsymbol{\theta}_l\right) p(l)} \right\}. \tag{40}$$

Page 637

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

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \boldsymbol{\Gamma}\right) \tag{41}$$

and so on.

Page 638

The first paragraph, Line 2: "conditional on" should read "conditioned on."

Page 641

The text after Equation (13.103): 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, the text should read for example: ..., we see that $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ is a Gaussian of the form

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = \mathcal{N}\left(\begin{pmatrix} \mathbf{z}_{n-1} \\ \mathbf{z}_n \end{pmatrix} \middle| \begin{pmatrix} \hat{\boldsymbol{\mu}}_{n-1} \\ \hat{\boldsymbol{\mu}}_n \end{pmatrix}, \begin{pmatrix} \hat{\mathbf{V}}_{n-1} & \hat{\mathbf{V}}_{n-1,n} \\ \hat{\mathbf{V}}_{n-1,n}^{\mathrm{T}} & \hat{\mathbf{V}}_n \end{pmatrix} \right) \tag{42}$$

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

$$\hat{\mathbf{V}}_{n-1,n} = \mathrm{cov}\left[\mathbf{z}_{n-1}, \mathbf{z}_n\right] = \mathbf{J}_{n-1}\hat{\mathbf{V}}_n. \tag{43}$$

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), Equation (13.109) should read

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \mathbb{E}_{\mathbf{Z}}\left[\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}\right)\middle|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right] \tag{44}$$

$$= \int \mathrm{d}\mathbf{Z}\, p\left(\mathbf{Z}\middle|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right) \ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}\right) \tag{45}$$

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

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 658

Figure 14.1, the equation below: 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 Equation (14.37) as, for example,

$$\gamma_{nk} = \mathbb{E}\left[z_{nk}\middle|t_n, \boldsymbol{\theta}^{\text{old}}\right] \tag{46}$$

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

$$\mathbb{E}\left[z_{nk}|t_n, \boldsymbol{\theta}\right] = p\left(z_{nk} = 1|t_n, \boldsymbol{\theta}\right) = \frac{\pi_k \mathcal{N}\left(t_n\middle|\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}\right)}{\sum_j \pi_j \mathcal{N}\left(t_n\middle|\mathbf{w}_j^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}\right)}. \tag{47}$$

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

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{t}, \mathbf{Z}|\boldsymbol{\theta})\Big|\mathbf{t}, \boldsymbol{\theta}^{\text{old}}\right] \tag{48}$$

$$= \ldots \tag{49}$$

where we have again used the typeface $\mathcal{Q}$ for the Q function so that the notation is consistent with that of Chapter 9.

Page 669

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

Page 669

Equation (14.41): $\boldsymbol{\Phi}$ is undefined. The text following (14.41) should read: ... where $\mathbf{R}_k = \text{diag}(\gamma_{nk})$ is a diagonal matrix of size $N \times N$ and $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N)^{\mathrm{T}}$ 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 Equation (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).

Page 671

Equation (14.48): The notation should be corrected similarly to the above erratum regarding (14.37).

Page 671

Equation (14.49): The notation should be corrected similarly to the above erratum regarding the unlabeled equation between (14.37) and (14.38).

Page 672

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

$$\nabla_k \nabla_k \mathcal{Q} = -\sum_{n=1}^{N} \gamma_{nk} y_{nk}(1 - y_{nk}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}. \tag{50}$$

Page 674

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

Page 686

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

Page 687

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

Page 688

Line 1: 'Gamma' should read 'gamma' (without capitalization).

Page 689

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.54): When identifying the functional form of this distribution in, e.g., the posterior distribution for the Gaussian mixture model of Section 9.2, I have found it helpful to write the distribution in terms of unnormalized probabilities $\tilde{\mu}_k \geq 0$ so that

$$\text{Mult}\left(\mathbf{x}|\tilde{\boldsymbol{\mu}}\right) = C(\tilde{\boldsymbol{\mu}}) \prod_{k=1}^{K} \tilde{\mu}_k^{x_k} \tag{51}$$

where we have defined the normalization constant as

$$C(\tilde{\boldsymbol{\mu}})^{-1} = \sum_{k=1}^{K} \tilde{\mu}_k. \tag{52}$$

and, thus, we have $\mu_k = C(\tilde{\boldsymbol{\mu}})\tilde{\mu}_k$.

Page 692

Equation (B.68): This form of multivariate Student's t-distribution is derived in Section 2.3.7 by marginalizing over a gamma distributed scalar variable η in (2.161), but *not* by marginalizing over the precision matrix that is generated from the conjugate Wishart distribution, which results in a distribution of the form

$$p\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{W}, \nu\right) = \int \mathcal{N}\left(\mathbf{x}\middle|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}|\mathbf{W}, \nu\right) \mathrm{d}\boldsymbol{\Lambda}. \tag{53}$$

I am not sure how we can marginalize over the precision matrix and, moreover, whether the marginalized distribution (53) is equivalent to (B.68). Some appropriate citation is needed if such details are omitted from the text.

Page 693

Equations (B.78) through (B.82): Some appropriate citation 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

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

$$\left(\mathbf{P}^{-1} + \mathbf{A}\mathbf{R}^{-1}\mathbf{B}\right)^{-1} \mathbf{A}\mathbf{R}^{-1} = \mathbf{P}\mathbf{A}\left(\mathbf{B}\mathbf{P}\mathbf{A} + \mathbf{R}\right)^{-1} \tag{54}$$

which is necessary to show the *push-through identity* (C.6) and also the determinant identity (C.14). As suggested in the text, the above identity (54) can be directly verified by right multiplying both sides by $(\mathbf{BPA} + \mathbf{R})$. However, I would prefer to prove the general push-through identity (54) 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

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \tag{55}$$

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

$$\begin{pmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{M}/\mathbf{A} \end{pmatrix} \tag{56}$$

or

$$\begin{pmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{M}/\mathbf{D} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{pmatrix} \tag{57}$$

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

$$\mathbf{M}/\mathbf{A} \equiv \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \tag{58}$$

or

$$\mathbf{M}/\mathbf{D} \equiv \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \tag{59}$$

respectively.[4] The above block diagonalization identities (56) and (57) yield two versions of the inverse partitioned matrix $\mathbf{M}^{-1}$, i.e.,

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{I} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{O} \\ \mathbf{O} & (\mathbf{M}/\mathbf{A})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \tag{62}$$

$$= \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{M}/\mathbf{A}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{M}/\mathbf{A}\right)^{-1} \\ -\left(\mathbf{M}/\mathbf{A}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & \left(\mathbf{M}/\mathbf{A}\right)^{-1} \end{pmatrix} \tag{63}$$

[4] 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 (56) and (57), we have from the definition of the determinant (C.10) that

$$\det(\mathbf{M}) = \det(\mathbf{A})\det(\mathbf{M}/\mathbf{A}) \tag{60}$$

$$\det(\mathbf{M}) = \det(\mathbf{D})\det(\mathbf{M}/\mathbf{D}). \tag{61}$$

and

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{I} & \mathbf{O}\\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I}\end{pmatrix}\begin{pmatrix}(\mathbf{M}/\mathbf{D})^{-1} & \mathbf{O}\\ \mathbf{O} & \mathbf{D}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{I} & -\mathbf{B}\mathbf{D}^{-1}\\ \mathbf{O} & \mathbf{I}\end{pmatrix} \tag{64}$$

$$= \begin{pmatrix}(\mathbf{M}/\mathbf{D})^{-1} & -(\mathbf{M}/\mathbf{D})^{-1}\mathbf{B}\mathbf{D}^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{M}/\mathbf{D})^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}(\mathbf{M}/\mathbf{D})^{-1}\mathbf{B}\mathbf{D}^{-1}\end{pmatrix} \tag{65}$$

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

$$(\mathbf{M}/\mathbf{D})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{M}/\mathbf{A})^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{66}$$

and

$$-(\mathbf{M}/\mathbf{A})^{-1}\mathbf{C}\mathbf{A}^{-1} = -\mathbf{D}^{-1}\mathbf{C}(\mathbf{M}/\mathbf{D})^{-1}. \tag{67}$$

Substituting (58) and (59) into both sides and replacing $\mathbf{D}$ with $-\mathbf{D}$, we finally have

$$\left(\mathbf{A}+\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D}+\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{68}$$

and

$$\left(\mathbf{D}+\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} = \mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A}+\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} \tag{69}$$

which are equivalent to (C.7) and (54), respectively.

Page 697

Equation (C.17): It is clear that the definition (C.17) of the derivative of a scalar with respect to a vector contradicts (C.16) and (C.18). The vector derivative of the form (C.17) is called the *gradient* whereas (C.18) is called the *Jacobian* (Minka, 2000). 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

$$\nabla_{\mathbf{x}}\mathbf{y} = \left(\frac{\partial y_j}{\partial x_i}\right) = \begin{pmatrix}\frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_M}{\partial x_1}\\ \vdots & \ddots & \vdots\\ \frac{\partial y_1}{\partial x_D} & \cdots & \frac{\partial y_M}{\partial x_D}\end{pmatrix}. \tag{70}$$

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.,

$$\nabla_{\mathbf{x}} y = \left(\frac{\partial y}{\partial x_i}\right) = \begin{pmatrix}\frac{\partial y}{\partial x_1}\\ \vdots\\ \frac{\partial y}{\partial x_D}\end{pmatrix}. \tag{71}$$

Note also that the right hand side of the definition of the gradient (70) 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.,

$$\nabla_{\mathbf{x}}\mathbf{z}(\mathbf{y}) = \left(\frac{\partial\mathbf{z}}{\partial\mathbf{y}}\frac{\partial\mathbf{y}}{\partial\mathbf{x}}\right)^{\mathrm{T}} = \nabla_{\mathbf{x}}\mathbf{y}\nabla_{\mathbf{y}}\mathbf{z}. \tag{72}$$

Since the chain rule (72) is handy when we compute the gradients of composite functions, I would suggest that it should also be pointed out in "(Vector and) Matrix Derivatives" of Appendix C.[5]

[5]At this point, one might wonder why we use 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 $\partial\mathbf{y}/\partial\mathbf{x}$. That is because Jacobians are useful in calculus while gradients are useful in optimization (Minka, 2000).

Page 697

Equation (C.19): Following the gradient notation (70), (C.19) should read

$$\nabla\left\{\mathbf{x}^{\mathrm{T}}\mathbf{a}\right\} = \nabla\left\{\mathbf{a}^{\mathrm{T}}\mathbf{x}\right\} = \mathbf{a} \tag{73}$$

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

$$\nabla\left\{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\right\} = \nabla\,\mathrm{Tr}\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{A}\right) = \left(\mathbf{A}^{\mathrm{T}} + \mathbf{A}\right)\mathbf{x} \tag{74}$$

$$\nabla\left\{\mathbf{B}\mathbf{x}\right\} = \mathbf{B}^{\mathrm{T}} \tag{75}$$

$$\nabla\left\{\varphi\mathbf{y}\right\} = \nabla\varphi\mathbf{y}^{\mathrm{T}} + \varphi\nabla\mathbf{y} \tag{76}$$

where matrices $\mathbf{A}$ and $\mathbf{B}$ are constants. Note that $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ in (74) 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

$$\nabla\left\{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\right\} = 2\mathbf{A}\mathbf{x}. \tag{77}$$

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

$$\nabla\left\|\mathbf{x}\right\|^2 = 2\mathbf{x} \tag{78}$$

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 (78) 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

$$E(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right\|^2. \tag{79}$$

Taking the gradient of (79) with respect to $\mathbf{w}$, we have

$$\nabla_{\mathbf{w}}E(\mathbf{w}) = -\mathbf{\Phi}^{\mathrm{T}}\left(\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right) \tag{80}$$

where we have used the identity (78) together with the chain rule (72) and the identity (75). The same result can also be obtained by first expanding the square norm in (79) and then differentiating it using the gradient identities given above. We use the identity (76) 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

$$J = \frac{1}{2}\sum_n \varepsilon_n^2. \tag{81}$$

The gradient and the Hessian of J can be evaluated as

$$\nabla J = \sum_n \varepsilon_n \nabla\varepsilon_n \tag{82}$$

$$\nabla\nabla J = \sum_n \nabla\varepsilon_n \left(\nabla\varepsilon_n\right)^{\mathrm{T}} + \sum_n \varepsilon_n \nabla\nabla\varepsilon_n. \tag{83}$$

It is also worth noting here that we can write down the Taylor series expansion (up to the second order) of a scalar function $f(\mathbf{x})$ compactly in terms of the gradients as

$$f(\mathbf{x} + \Delta\mathbf{x}) \simeq f(\mathbf{x}) + \mathbf{g}^{\mathrm{T}}\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^{\mathrm{T}}\mathbf{H}\Delta\mathbf{x} \tag{84}$$

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

$$\mathbf{g} \equiv \nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_D} \end{pmatrix}, \quad \mathbf{H} \equiv \nabla\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_D \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_D \partial x_D} \end{pmatrix}. \tag{85}$$

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 is a matrix with the same dimensionality as $\mathbf{A}$ and is given by

$$\frac{\partial \mathbf{A}}{\partial x} = \left(\frac{\partial A_{ij}}{\partial x} \right) \tag{86}$$

which is analogous to (C.18) in that the partial derivatives are laid out according to the numerator, i.e., $\mathbf{A}$ in (86). On the other hand, the gradient (70) is such that the derivatives are laid out according to the denominator. We should also point out that, in a similar analogy, we write the gradient $\nabla_{\mathbf{A}} y$ of a scalar y with respect to a matrix $\mathbf{A}$ as

$$\nabla_{\mathbf{A}} y = \left(\frac{\partial y}{\partial A_{ij}} \right). \tag{87}$$

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 notation (4) for $|\mathbf{A}|$, which I would recommend, we see that (C.22) holds for any nonsingular $\mathbf{A}$ such that $\det(\mathbf{A}) \neq 0$. 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. To this end, we first show the following identity for any square matrix $\mathbf{A}$

$$\frac{\partial}{\partial x} \det(\mathbf{A}) = \mathrm{Tr}\left(\mathbf{A}^{\dagger} \frac{\partial \mathbf{A}}{\partial x} \right) \tag{88}$$

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

$$A^{\dagger}_{ij} = (-1)^{i+j} \det\left(\mathbf{A}^{(ji)} \right) \tag{89}$$

where $\mathbf{A}^{(ij)}$ is a matrix obtained by removing the i-th row and the j-th column of $\mathbf{A}$. From the identity

$$\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}\mathbf{A} = \det(\mathbf{A})\mathbf{I} \tag{90}$$

we can write the inverse matrix $\mathbf{A}^{-1}$ in terms of the adjugate matrix $\mathbf{A}^{\dagger}$ so that

$$\mathbf{A}^{-1} = \frac{\mathbf{A}^{\dagger}}{\det(\mathbf{A})} \tag{91}$$

if $\mathbf{A}$ is nonsingular so that $\det(\mathbf{A}) \neq 0$. Note also that the above identity (90) implies

$$\det(\mathbf{A}) = \sum_k A_{ik} A^{\dagger}_{ki} = \sum_k A^{\dagger}_{jk} A_{kj} \tag{92}$$

for any i and j. Substituting this identity (92) into the left hand side of (88) and noting that, from the definition (89) of the adjugate matrix, $A^{\dagger}_{ji}$ is independent of A_{ik} nor A_{kj} for any k, we

have

$$\frac{\partial}{\partial x}\det(\mathbf{A}) = \sum_{ij}\left\{\frac{\partial}{\partial A_{ij}}\sum_k A_{ik}A_{ki}^{\dagger}\right\}\frac{\partial A_{ij}}{\partial x} = \sum_{ij}\left\{\frac{\partial}{\partial A_{ij}}\sum_k A_{jk}^{\dagger}A_{kj}\right\}\frac{\partial A_{ij}}{\partial x} \tag{93}$$

$$= \sum_{ij} A_{ji}^{\dagger}\frac{\partial A_{ij}}{\partial x} \tag{94}$$

$$= \mathrm{Tr}\left(\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{\dagger}\right) = \mathrm{Tr}\left(\mathbf{A}^{\dagger}\frac{\partial \mathbf{A}}{\partial x}\right) \tag{95}$$

which proves the identity (88). Making use of (88) together with (91), we can now evaluate the right hand side of (C.22) as

$$\frac{\partial}{\partial x}\ln|\mathbf{A}| = \frac{1}{\det(\mathbf{A})}\frac{\partial}{\partial x}\det(\mathbf{A}) \tag{96}$$

$$= \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right) \tag{97}$$

if $\mathbf{A}$ is nonsingular so that $\det(\mathbf{A}) \neq 0$ where we have used the notation (4) for $|\mathbf{A}|$.

Page 698

Equations (C.24), (C.25), (C.26), (C.27), and (C.28): Since these derivatives are gradients, the operator $\frac{\partial}{\partial \mathbf{A}}$ should read $\nabla_{\mathbf{A}}$ if we adopt the notation (87) for the gradient of a scalar with respect to a matrix. For example, (C.28) should read

$$\nabla_{\mathbf{A}}\ln|\mathbf{A}| = \mathbf{A}^{-\mathrm{T}} \tag{98}$$

where we have used (C.4) and written

$$\left(\mathbf{A}^{-1}\right)^{\mathrm{T}} = \left(\mathbf{A}^{\mathrm{T}}\right)^{-1} = \mathbf{A}^{-\mathrm{T}}. \tag{99}$$

In addition to these identities regarding gradients with respect to a matrix, I would suggest to include two more identities

$$\nabla_{\mathbf{A}}\mathrm{Tr}\left(\mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{C}\right) = \mathbf{C}^{\mathrm{T}}\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{C}\mathbf{A}\mathbf{B} \tag{100}$$

$$\nabla_{\mathbf{A}}\mathrm{Tr}\left(\mathbf{A}^{-1}\mathbf{B}\right) = -\mathbf{A}^{-\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-\mathrm{T}}. \tag{101}$$

We use the above identities (100) and (101), 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 the former identity (100).

Page 717

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

References

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

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

Minka, T. P. (2000). Old and new matrix algebra useful for statistics. http://research.microsoft.com/en-us/um/people/minka/papers/matrix/minka-matrix.pdf.

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.

Svensén, M. and C. M. Bishop (2009). Pattern recognition and machine learning: Solutions to the exercises, web edition. http://research.microsoft.com/en-us/um/people/cmbishop/prml/pdf/prml-web-sol-2009-09-08.pdf.

Svensén, M. and C. M. Bishop (2011). Pattern recognition and machine learning: Errata and additional comments. http://research.microsoft.com/en-us/um/people/cmbishop/prml/pdf/prml-errata-2nd-20110921.pdf.