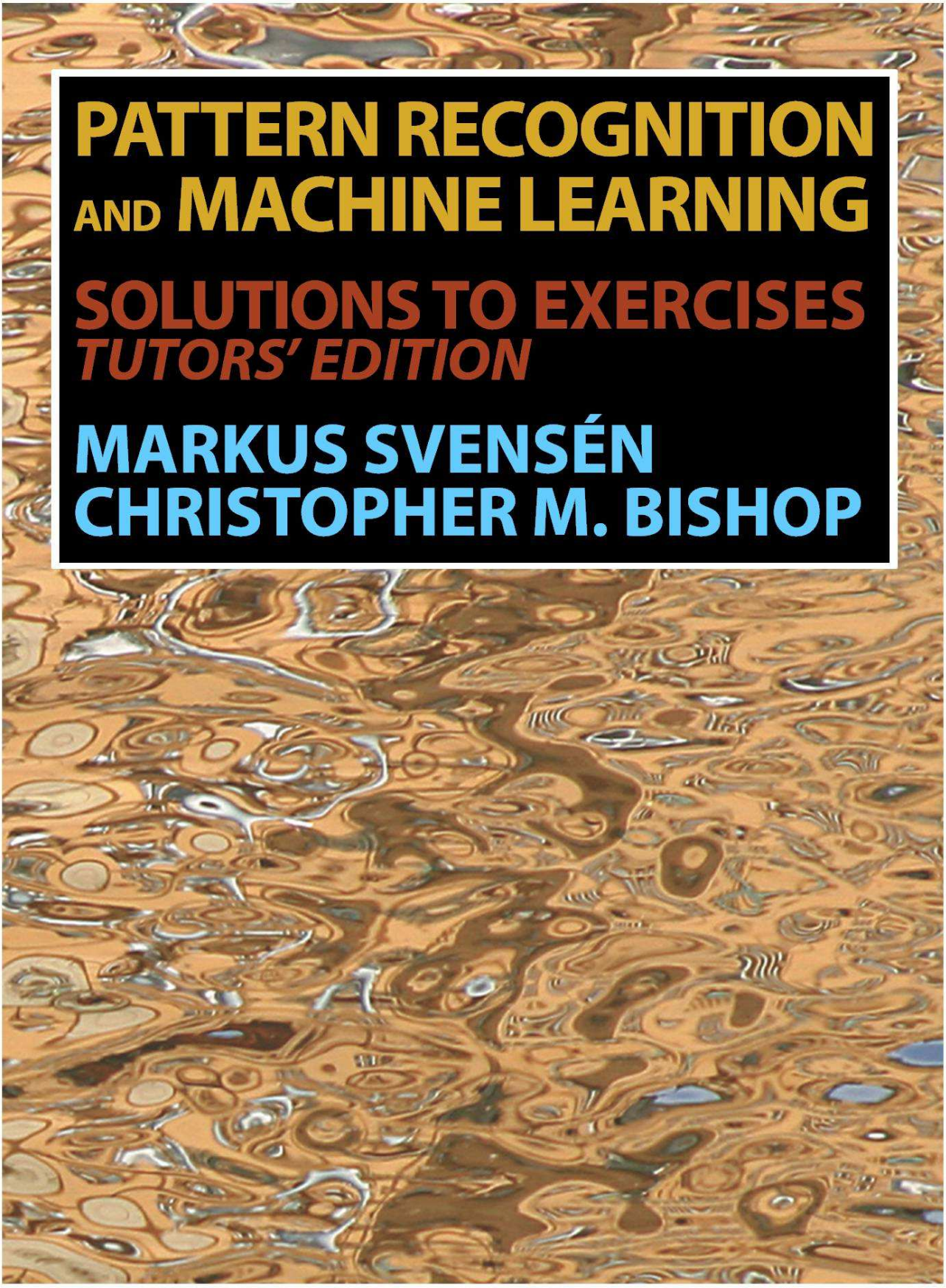


PATTERN RECOGNITION AND MACHINE LEARNING

SOLUTIONS TO EXERCISES
TUTORS' EDITION

MARKUS SVENSÉN
CHRISTOPHER M. BISHOP

Pattern Recognition and Machine Learning

Solutions to the Exercises: Tutors' Edition

Markus Svensén and Christopher M. Bishop

Copyright © 2002–2009

This is the solutions manual (Tutors' Edition) for the book *Pattern Recognition and Machine Learning* (PRML; published by Springer in 2006). This release was created September 8, 2009. Any future releases (e.g. with corrections to errors) will be announced on the PRML web-site (see below) and published via Springer.

PLEASE DO NOT DISTRIBUTE

Most of the solutions in this manual are intended as a resource for tutors teaching courses based on PRML and the value of this resource would be greatly diminished if was to become generally available. All tutors who want a copy should contact Springer directly.

The authors would like to express their gratitude to the various people who have provided feedback on earlier releases of this document.

The authors welcome all comments, questions and suggestions about the solutions as well as reports on (potential) errors in text or formulae in this document; please send any such feedback to

`prml-fb@microsoft.com`

Further information about PRML is available from

http://research.microsoft.com/∼cmbishop/PRML

Contents

Contents **5**
Chapter 1: Introduction 7
Chapter 2: Probability Distributions 28
Chapter 3: Linear Models for Regression 62
Chapter 4: Linear Models for Classification 78
Chapter 5: Neural Networks 93
Chapter 6: Kernel Methods 114
Chapter 7: Sparse Kernel Machines 128
Chapter 8: Graphical Models 136
Chapter 9: Mixture Models and EM 150
Chapter 10: Approximate Inference 163
Chapter 11: Sampling Methods 198
Chapter 12: Continuous Latent Variables 207
Chapter 13: Sequential Data 223
Chapter 14: Combining Models 246

Chapter 1 Introduction

1.1 Substituting (1.1) into (1.2) and then differentiating with respect to w_i we obtain

$$\sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x_n^j - t_n\right) x_n^i = 0. \tag{1}$$

Re-arranging terms then gives the required result.

1.2 For the regularized sum-of-squares error function given by (1.4) the corresponding linear equations are again obtained by differentiation, and take the same form as (1.122), but with A_{ij} replaced by $\widetilde{A}_{ij}$, given by

$$\widetilde{A}_{ij} = A_{ij} + \lambda I_{ij}. \tag{2}$$

1.3 Let us denote apples, oranges and limes by a, o and l respectively. The marginal probability of selecting an apple is given by

$$\begin{aligned} p(a) &= p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g) \\ &= \frac{3}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.34 \end{aligned} \tag{3}$$

where the conditional probabilities are obtained from the proportions of apples in each box.

To find the probability that the box was green, given that the fruit we selected was an orange, we can use Bayes' theorem

$$p(g|o) = \frac{p(o|g)p(g)}{p(o)}. \tag{4}$$

The denominator in (4) is given by

$$\begin{aligned} p(o) &= p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g) \\ &= \frac{4}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.36 \end{aligned} \tag{5}$$

from which we obtain

$$p(g|o) = \frac{3}{10} \times \frac{0.6}{0.36} = \frac{1}{2}. \tag{6}$$

1.4 We are often interested in finding the most probable value for some quantity. In the case of probability distributions over discrete variables this poses little problem. However, for continuous variables there is a subtlety arising from the nature of probability densities and the way they transform under non-linear changes of variable.

Consider first the way a function $f(x)$ behaves when we change to a new variable y where the two variables are related by $x = g(y)$. This defines a new function of y given by

$$\widetilde{f}(y) = f(g(y)). \tag{7}$$

Suppose $f(x)$ has a mode (i.e. a maximum) at $\widehat{x}$ so that $f'(\widehat{x}) = 0$. The corresponding mode of $\widetilde{f}(y)$ will occur for a value $\widehat{y}$ obtained by differentiating both sides of (7) with respect to y

$$\widetilde{f}'(\widehat{y}) = f'(g(\widehat{y}))g'(\widehat{y}) = 0. \tag{8}$$

Assuming $g'(\widehat{y}) \neq 0$ at the mode, then $f'(g(\widehat{y})) = 0$. However, we know that $f'(\widehat{x}) = 0$, and so we see that the locations of the mode expressed in terms of each of the variables x and y are related by $\widehat{x} = g(\widehat{y})$, as one would expect. Thus, finding a mode with respect to the variable x is completely equivalent to first transforming to the variable y, then finding a mode with respect to y, and then transforming back to x.

Now consider the behaviour of a probability density $p_x(x)$ under the change of variables $x = g(y)$, where the density with respect to the new variable is $p_y(y)$ and is given by ((1.27)). Let us write $g'(y) = s|g'(y)|$ where $s \in \{-1, +1\}$. Then ((1.27)) can be written

$$p_y(y) = p_x(g(y))sg'(y).$$

Differentiating both sides with respect to y then gives

$$p'_y(y) = sp'_x(g(y))\{g'(y)\}^2 + sp_x(g(y))g''(y). \tag{9}$$

Due to the presence of the second term on the right hand side of (9) the relationship $\widehat{x} = g(\widehat{y})$ no longer holds. Thus the value of x obtained by maximizing $p_x(x)$ will not be the value obtained by transforming to $p_y(y)$ then maximizing with respect to y and then transforming back to x. This causes modes of densities to be dependent on the choice of variables. In the case of linear transformation, the second term on the right hand side of (9) vanishes, and so the location of the maximum transforms according to $\widehat{x} = g(\widehat{y})$.

This effect can be illustrated with a simple example, as shown in Figure 1. We begin by considering a Gaussian distribution $p_x(x)$ over x with mean $\mu = 6$ and standard deviation $\sigma = 1$, shown by the red curve in Figure 1. Next we draw a sample of $N = 50,000$ points from this distribution and plot a histogram of their values, which as expected agrees with the distribution $p_x(x)$.

Now consider a non-linear change of variables from x to y given by

$$x = g(y) = \ln(y) - \ln(1 - y) + 5. \tag{10}$$

The inverse of this function is given by

$$y = g^{-1}(x) = \frac{1}{1 + \exp(-x + 5)} \tag{11}$$

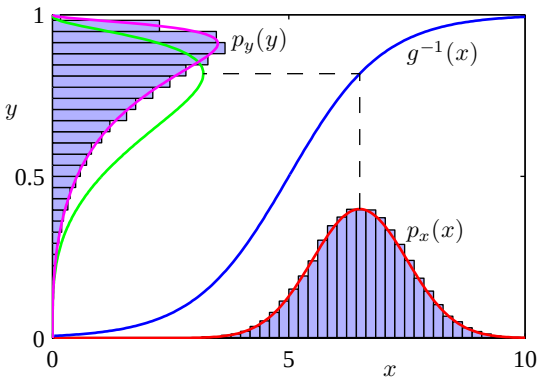


Figure 1 Example of the transformation of the mode of a density under a non-linear change of variables, illustrating the different behaviour compared to a simple function. See the text for details.

which is a *logistic sigmoid* function, and is shown in Figure 1 by the blue curve.

If we simply transform $p_x(x)$ as a function of x we obtain the green curve $p_x(g(y))$ shown in Figure 1, and we see that the mode of the density $p_x(x)$ is transformed via the sigmoid function to the mode of this curve. However, the density over y transforms instead according to (1.27) and is shown by the magenta curve on the left side of the diagram. Note that this has its mode shifted relative to the mode of the green curve.

To confirm this result we take our sample of $50,000$ values of x, evaluate the corresponding values of y using (11), and then plot a histogram of their values. We see that this histogram matches the magenta curve in Figure 1 and not the green curve!

1.5 Expanding the square we have

$$\begin{aligned}\mathbb{E}[(f(x)-\mathbb{E}[f(x)])^2] &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2\end{aligned}$$

as required.

1.6 The definition of covariance is given by (1.41) as

$$\mathrm{cov}[x,y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y].$$

Using (1.33) and the fact that $p(x,y) = p(x)p(y)$ when x and y are independent, we obtain

$$\begin{aligned}\mathbb{E}[xy] &= \sum_x \sum_y p(x,y)xy \\ &= \sum_x p(x)x \sum_y p(y)y \\ &= \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

and hence $\mathrm{cov}[x, y] = 0$. The case where x and y are continuous variables is analogous, with (1.33) replaced by (1.34) and the sums replaced by integrals.

1.7 The transformation from Cartesian to polar coordinates is defined by

$$x = r\cos\theta \tag{12}$$
$$y = r\sin\theta \tag{13}$$

and hence we have $x^2 + y^2 = r^2$ where we have used the well-known trigonometric result (2.177). Also the Jacobian of the change of variables is easily seen to be

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

where again we have used (2.177). Thus the double integral in (1.125) becomes

$$I^2 = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) r\,\mathrm{d}r\,\mathrm{d}\theta \tag{14}$$
$$= 2\pi\int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2}\right)\frac{1}{2}\,\mathrm{d}u \tag{15}$$
$$= \pi\left[\exp\left(-\frac{u}{2\sigma^2}\right)\left(-2\sigma^2\right)\right]_0^{\infty} \tag{16}$$
$$= 2\pi\sigma^2 \tag{17}$$

where we have used the change of variables $r^2 = u$. Thus

$$I = \left(2\pi\sigma^2\right)^{1/2}.$$

Finally, using the transformation $y = x - \mu$, the integral of the Gaussian distribution becomes

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\int_{-\infty}^{\infty}\exp\left(-\frac{y^2}{2\sigma^2}\right)\,\mathrm{d}y$$
$$= \frac{I}{\left(2\pi\sigma^2\right)^{1/2}} = 1$$

as required.

1.8 From the definition (1.46) of the univariate Gaussian distribution, we have

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}x\,\mathrm{d}x. \tag{18}$$

Now change variables using $y = x - \mu$ to give

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}y^2\right\} (y + \mu)\, \mathrm{d}y. \tag{19}$$

We now note that in the factor $(y + \mu)$ the first term in y corresponds to an odd integrand and so this integral must vanish (to show this explicitly, write the integral as the sum of two integrals, one from $-\infty$ to 0 and the other from 0 to ∞ and then show that these two integrals cancel). In the second term, μ is a constant and pulls outside the integral, leaving a normalized Gaussian distribution which integrates to 1, and so we obtain (1.49).

To derive (1.50) we first substitute the expression (1.46) for the normal distribution into the normalization result (1.48) and re-arrange to obtain

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \mathrm{d}x = \left(2\pi\sigma^2\right)^{1/2}. \tag{20}$$

We now differentiate both sides of (20) with respect to σ^2 and then re-arrange to obtain

$$\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} (x - \mu)^2\, \mathrm{d}x = \sigma^2 \tag{21}$$

which directly shows that

$$\mathbb{E}[(x - \mu)^2] = \mathrm{var}[x] = \sigma^2. \tag{22}$$

Now we expand the square on the left-hand side giving

$$\mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 = \sigma^2.$$

Making use of (1.49) then gives (1.50) as required.

Finally, (1.51) follows directly from (1.49) and (1.50)

$$\mathbb{E}[x^2] - \mathbb{E}[x]^2 = \left(\mu^2 + \sigma^2\right) - \mu^2 = \sigma^2.$$

1.9 For the univariate case, we simply differentiate (1.46) with respect to x to obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}\left(x|\mu, \sigma^2\right) = -\mathcal{N}\left(x|\mu, \sigma^2\right)\frac{x - \mu}{\sigma^2}.$$

Setting this to zero we obtain $x = \mu$.

Similarly, for the multivariate case we differentiate (1.52) with respect to $\mathbf{x}$ to obtain

$$\begin{aligned}\frac{\partial}{\partial\mathbf{x}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{1}{2}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\nabla_{\mathbf{x}}\left\{(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \\ &= -\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}),\end{aligned}$$

where we have used (C.19), (C.20)[1] and the fact that $\boldsymbol{\Sigma}^{-1}$ is symmetric. Setting this derivative equal to $\mathbf{0}$, and left-multiplying by $\boldsymbol{\Sigma}$, leads to the solution $\mathbf{x} = \boldsymbol{\mu}$.

[1] **NOTE**: In the 1st printing of PRML, there are mistakes in (C.20); all instances of $\mathbf{x}$ (vector) in the denominators should be x (scalar).

1.10 Since x and z are independent, their joint distribution factorizes $p(x, z) = p(x)p(z)$, and so

$$\begin{aligned}\mathbb{E}[x+z] &= \iint (x+z)p(x)p(z)\,\mathrm{d}x\,\mathrm{d}z &(23)\\ &= \int xp(x)\,\mathrm{d}x + \int zp(z)\,\mathrm{d}z &(24)\\ &= \mathbb{E}[x] + \mathbb{E}[z]. &(25)\end{aligned}$$

Similarly for the variances, we first note that

$$(x+z-\mathbb{E}[x+z])^2 = (x-\mathbb{E}[x])^2 + (z-\mathbb{E}[z])^2 + 2(x-\mathbb{E}[x])(z-\mathbb{E}[z]) \quad (26)$$

where the final term will integrate to zero with respect to the factorized distribution $p(x)p(z)$. Hence

$$\begin{aligned}\mathrm{var}[x+z] &= \iint (x+z-\mathbb{E}[x+z])^2 p(x)p(z)\,\mathrm{d}x\,\mathrm{d}z\\ &= \int (x-\mathbb{E}[x])^2 p(x)\,\mathrm{d}x + \int (z-\mathbb{E}[z])^2 p(z)\,\mathrm{d}z\\ &= \mathrm{var}(x) + \mathrm{var}(z). &(27)\end{aligned}$$

For discrete variables the integrals are replaced by summations, and the same results are again obtained.

1.11 We use ℓ to denote $\ln p(\mathbf{X}|\mu, \sigma^2)$ from (1.54). By standard rules of differentiation we obtain

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu).$$

Setting this equal to zero and moving the terms involving μ to the other side of the equation we get

$$\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n = \frac{1}{\sigma^2} N\mu$$

and by multiplying ing both sides by σ^2/N we get (1.55).

Similarly we have

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\frac{1}{\sigma^2}$$

and setting this to zero we obtain

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

Multiplying both sides by $2(\sigma^2)^2/N$ and substituting μ_{ML} for μ we get (1.56).

1.12 If $m = n$ then $x_n x_m = x_n^2$ and using (1.50) we obtain $\mathbb{E}[x_n^2] = \mu^2 + \sigma^2$, whereas if $n \neq m$ then the two data points x_n and x_m are independent and hence $\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2$ where we have used (1.49). Combining these two results we obtain (1.130).

Next we have

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] = \mu \tag{28}$$

using (1.49).

Finally, consider $\mathbb{E}[\sigma^2_{\mathrm{ML}}]$. From (1.55) and (1.56), and making use of (1.130), we have

$$\begin{aligned}
\mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(x_n - \frac{1}{N}\sum_{m=1}^{N}x_m\right)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - \frac{2}{N}x_n\sum_{m=1}^{N}x_m + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}x_m x_l\right] \\
&= \left\{\mu^2 + \sigma^2 - 2\left(\mu^2 + \frac{1}{N}\sigma^2\right) + \mu^2 + \frac{1}{N}\sigma^2\right\} \\
&= \left(\frac{N-1}{N}\right)\sigma^2
\end{aligned} \tag{29}$$

as required.

1.13 In a similar fashion to solution 1.12, substituting μ for μ_{ML} in (1.56) and using (1.49) and (1.50) we have

$$\begin{aligned}
\mathbb{E}_{\{\mathbf{x}_n\}}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}_n}\left[x_n^2 - 2x_n\mu + \mu^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - 2\mu\mu + \mu^2\right) \\
&= \sigma^2
\end{aligned}$$

1.14 Define

$$w_{ij}^{\mathrm{S}} = \frac{1}{2}(w_{ij} + w_{ji}) \qquad w_{ij}^{\mathrm{A}} = \frac{1}{2}(w_{ij} - w_{ji}). \tag{30}$$

from which the (anti)symmetry properties follow directly, as does the relation $w_{ij} = w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}}$. We now note that

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_i x_j = \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}x_i x_j - \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ji}x_i x_j = 0 \tag{31}$$

from which we obtain (1.132). The number of independent components in w_{ij}^{S} can be found by noting that there are D^2 parameters in total in this matrix, and that entries off the leading diagonal occur in constrained pairs $w_{ij} = w_{ji}$ for $j \neq i$. Thus we start with D^2 parameters in the matrix w_{ij}^{S}, subtract D for the number of parameters on the leading diagonal, divide by two, and then add back D for the leading diagonal and we obtain $(D^2 - D)/2 + D = D(D+1)/2$.

1.15 The redundancy in the coefficients in (1.133) arises from interchange symmetries between the indices i_k. Such symmetries can therefore be removed by enforcing an ordering on the indices, as in (1.134), so that only one member in each group of equivalent configurations occurs in the summation.

To derive (1.135) we note that the number of independent parameters $n(D, M)$ which appear at order M can be written as

$$n(D, M) = \sum_{i_1=1}^{D} \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 \tag{32}$$

which has M terms. This can clearly also be written as

$$n(D, M) = \sum_{i_1=1}^{D} \left\{ \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 \right\} \tag{33}$$

where the term in braces has $M-1$ terms which, from (32), must equal $n(i_1, M-1)$. Thus we can write

$$n(D, M) = \sum_{i_1=1}^{D} n(i_1, M-1) \tag{34}$$

which is equivalent to (1.135).

To prove (1.136) we first set $D = 1$ on both sides of the equation, and make use of $0! = 1$, which gives the value 1 on both sides, thus showing the equation is valid for $D = 1$. Now we assume that it is true for a specific value of dimensionality D and then show that it must be true for dimensionality $D+1$. Thus consider the left-hand side of (1.136) evaluated for $D+1$ which gives

$$\begin{aligned} \sum_{i=1}^{D+1} \frac{(i+M-2)!}{(i-1)!(M-1)!} &= \frac{(D+M-1)!}{(D-1)!M!} + \frac{(D+M-1)!}{D!(M-1)!} \\ &= \frac{(D+M-1)!D + (D+M-1)!M}{D!M!} \\ &= \frac{(D+M)!}{D!M!} \end{aligned} \tag{35}$$

which equals the right hand side of (1.136) for dimensionality $D+1$. Thus, by induction, (1.136) must hold true for all values of D.

Finally we use induction to prove (1.137). For $M = 2$ we find obtain the standard result $n(D, 2) = \frac{1}{2}D(D + 1)$, which is also proved in Exercise 1.14. Now assume that (1.137) is correct for a specific order $M - 1$ so that

$$n(D, M - 1) = \frac{(D + M - 2)!}{(D - 1)!\,(M - 1)!}. \tag{36}$$

Substituting this into the right hand side of (1.135) we obtain

$$n(D, M) = \sum_{i=1}^{D} \frac{(i + M - 2)!}{(i - 1)!\,(M - 1)!} \tag{37}$$

which, making use of (1.136), gives

$$n(D, M) = \frac{(D + M - 1)!}{(D - 1)!\,M!} \tag{38}$$

and hence shows that (1.137) is true for polynomials of order M. Thus by induction (1.137) must be true for all values of M.

1.16 **NOTE**: In the 1st printing of PRML, this exercise contains two typographical errors. On line 4, $M6th$ should be M^{th} and on the l.h.s. of (1.139), $N(d, M)$ should be $N(D, M)$.

The result (1.138) follows simply from summing up the coefficients at all order up to and including order M. To prove (1.139), we first note that when $M = 0$ the right hand side of (1.139) equals 1, which we know to be correct since this is the number of parameters at zeroth order which is just the constant offset in the polynomial. Assuming that (1.139) is correct at order M, we obtain the following result at order $M + 1$

$$\begin{aligned} N(D, M + 1) &= \sum_{m=0}^{M+1} n(D, m) \\ &= \sum_{m=0}^{M} n(D, m) + n(D, M + 1) \\ &= \frac{(D + M)!}{D!M!} + \frac{(D + M)!}{(D - 1)!(M + 1)!} \\ &= \frac{(D + M)!(M + 1) + (D + M)!D}{D!(M + 1)!} \\ &= \frac{(D + M + 1)!}{D!(M + 1)!} \end{aligned}$$

which is the required result at order $M + 1$.

Now assume $M \gg D$. Using Stirling's formula we have

$$\begin{aligned} n(D,M) &\simeq \frac{(D+M)^{D+M}e^{-D-M}}{D!\,M^M e^{-M}} \\ &= \frac{M^{D+M}e^{-D}}{D!\,M^M}\left(1+\frac{D}{M}\right)^{D+M} \\ &\simeq \frac{M^D e^{-D}}{D!}\left(1+\frac{D(D+M)}{M}\right) \\ &\simeq \frac{(1+D)e^{-D}}{D!}M^D \end{aligned}$$

which grows like M^D with M. The case where $D \gg M$ is identical, with the roles of D and M exchanged. By numerical evaluation we obtain $N(10,3) = 286$ and $N(100,3) = 176{,}851$.

1.17 Using integration by parts we have

$$\begin{aligned} \Gamma(x+1) &= \int_0^\infty u^x e^{-u}\,\mathrm{d}u \\ &= \left[-e^{-u}u^x\right]_0^\infty + \int_0^\infty xu^{x-1}e^{-u}\,\mathrm{d}u = 0 + x\Gamma(x). \end{aligned} \tag{39}$$

For $x = 1$ we have

$$\Gamma(1) = \int_0^\infty e^{-u}\,\mathrm{d}u = \left[-e^{-u}\right]_0^\infty = 1. \tag{40}$$

If x is an integer we can apply proof by induction to relate the gamma function to the factorial function. Suppose that $\Gamma(x+1) = x!$ holds. Then from the result (39) we have $\Gamma(x+2) = (x+1)\Gamma(x+1) = (x+1)!$. Finally, $\Gamma(1) = 1 = 0!$, which completes the proof by induction.

1.18 On the right-hand side of (1.142) we make the change of variables $u = r^2$ to give

$$\frac{1}{2}S_D\int_0^\infty e^{-u}u^{D/2-1}\,\mathrm{d}u = \frac{1}{2}S_D\Gamma(D/2) \tag{41}$$

where we have used the definition (1.141) of the Gamma function. On the left hand side of (1.142) we can use (1.126) to obtain $\pi^{D/2}$. Equating these we obtain the desired result (1.143).

The volume of a sphere of radius 1 in D-dimensions is obtained by integration

$$V_D = S_D\int_0^1 r^{D-1}\,\mathrm{d}r = \frac{S_D}{D}. \tag{42}$$

For $D = 2$ and $D = 3$ we obtain the following results

$$S_2 = 2\pi, \qquad S_3 = 4\pi, \qquad V_2 = \pi a^2, \qquad V_3 = \frac{4}{3}\pi a^3. \tag{43}$$

1.19 The volume of the cube is $(2a)^D$. Combining this with (1.143) and (1.144) we obtain (1.145). Using Stirling's formula (1.146) in (1.145) the ratio becomes, for large D,

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \left(\frac{\pi e}{2D}\right)^{D/2} \frac{1}{D} \tag{44}$$

which goes to 0 as $D \to \infty$. The distance from the center of the cube to the mid point of one of the sides is a, since this is where it makes contact with the sphere. Similarly the distance to one of the corners is $a\sqrt{D}$ from Pythagoras' theorem. Thus the ratio is $\sqrt{D}$.

1.20 Since $p(\mathbf{x})$ is radially symmetric it will be roughly constant over the shell of radius r and thickness ϵ. This shell has volume $S_D r^{D-1}\epsilon$ and since $\|\mathbf{x}\|^2 = r^2$ we have

$$\int_{\text{shell}} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \simeq p(r) S_D r^{D-1}\epsilon \tag{45}$$

from which we obtain (1.148). We can find the stationary points of $p(r)$ by differentiation

$$\frac{\mathrm{d}}{\mathrm{d}r}p(r) \propto \left[(D-1)r^{D-2} + r^{D-1}\left(-\frac{r}{\sigma^2}\right)\right] \exp\left(-\frac{r^2}{2\sigma^2}\right) = 0. \tag{46}$$

Solving for r, and using $D \gg 1$, we obtain $\widehat{r} \simeq \sqrt{D}\sigma$.

Next we note that

$$\begin{aligned} p(\widehat{r}+\epsilon) &\propto (\widehat{r}+\epsilon)^{D-1} \exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2} + (D-1)\ln(\widehat{r}+\epsilon)\right]. \end{aligned} \tag{47}$$

We now expand $p(r)$ around the point $\widehat{r}$. Since this is a stationary point of $p(r)$ we must keep terms up to second order. Making use of the expansion $\ln(1+x) = x - x^2/2 + O(x^3)$, together with $D \gg 1$, we obtain (1.149).

Finally, from (1.147) we see that the probability density at the origin is given by

$$p(\mathbf{x}=\mathbf{0}) = \frac{1}{(2\pi\sigma^2)^{1/2}}$$

while the density at $\|\mathbf{x}\| = \widehat{r}$ is given from (1.147) by

$$p(\|\mathbf{x}\| = \widehat{r}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\widehat{r}^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{D}{2}\right)$$

where we have used $\widehat{r} \simeq \sqrt{D}\sigma$. Thus the ratio of densities is given by $\exp(D/2)$.

1.21 Since the square root function is monotonic for non-negative numbers, we can take the square root of the relation $a \leqslant b$ to obtain $a^{1/2} \leqslant b^{1/2}$. Then we multiply both sides by the non-negative quantity $a^{1/2}$ to obtain $a \leqslant (ab)^{1/2}$.

The probability of a misclassification is given, from (1.78), by

$$\begin{aligned} p(\text{mistake}) &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{R}_1} p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathcal{C}_1|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \end{aligned} \tag{48}$$

Since we have chosen the decision regions to minimize the probability of misclassification we must have $p(\mathcal{C}_2|\mathbf{x}) \leqslant p(\mathcal{C}_1|\mathbf{x})$ in region $\mathcal{R}_1$, and $p(\mathcal{C}_1|\mathbf{x}) \leqslant p(\mathcal{C}_2|\mathbf{x})$ in region $\mathcal{R}_2$. We now apply the result $a \leqslant b \Rightarrow a^{1/2} \leqslant b^{1/2}$ to give

$$\begin{aligned} p(\text{mistake}) &\leqslant \int_{\mathcal{R}_1} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\quad + \int_{\mathcal{R}_2} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int \{p(\mathcal{C}_1|\mathbf{x})p(\mathbf{x})p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})\}^{1/2}\,\mathrm{d}\mathbf{x} \end{aligned} \tag{49}$$

since the two integrals have the same integrand. The final integral is taken over the whole of the domain of $\mathbf{x}$.

1.22 Substituting $L_{kj} = 1 - \delta_{kj}$ into (1.81), and using the fact that the posterior probabilities sum to one, we find that, for each $\mathbf{x}$ we should choose the class j for which $1 - p(\mathcal{C}_j|\mathbf{x})$ is a minimum, which is equivalent to choosing the j for which the posterior probability $p(\mathcal{C}_j|\mathbf{x})$ is a maximum. This loss matrix assigns a loss of one if the example is misclassified, and a loss of zero if it is correctly classified, and hence minimizing the expected loss will minimize the misclassification rate.

1.23 From (1.81) we see that for a general loss matrix and arbitrary class priors, the expected loss is minimized by assigning an input $\mathbf{x}$ to class the j which minimizes

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_k L_{kj} p(\mathbf{x}|\mathcal{C}_k) p(\mathcal{C}_k)$$

and so there is a direct trade-off between the priors $p(\mathcal{C}_k)$ and the loss matrix L_{kj}.

1.24 A vector $\mathbf{x}$ belongs to class $\mathcal{C}_k$ with probability $p(\mathcal{C}_k|\mathbf{x})$. If we decide to assign $\mathbf{x}$ to class $\mathcal{C}_j$ we will incur an expected loss of $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$, whereas if we select the reject option we will incur a loss of λ. Thus, if

$$j = \arg\min_l \sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) \tag{50}$$

then we minimize the expected loss if we take the following action

$$\text{choose} \begin{cases} \text{class } j, & \text{if } \min_l \sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) < \lambda; \\ \text{reject}, & \text{otherwise.} \end{cases} \tag{51}$$

For a loss matrix $L_{kj} = 1 - I_{kj}$ we have $\sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) = 1 - p(\mathcal{C}_l|\mathbf{x})$ and so we reject unless the smallest value of $1 - p(\mathcal{C}_l|\mathbf{x})$ is less than λ, or equivalently if the largest value of $p(\mathcal{C}_l|\mathbf{x})$ is less than $1 - \lambda$. In the standard reject criterion we reject if the largest posterior probability is less than θ. Thus these two criteria for rejection are equivalent provided $\theta = 1 - \lambda$.

1.25 The expected squared loss for a vectorial target variable is given by

$$\mathbb{E}[L] = \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{t}.$$

Our goal is to choose $\mathbf{y}(\mathbf{x})$ so as to minimize $\mathbb{E}[L]$. We can do this formally using the calculus of variations to give

$$\frac{\delta \mathbb{E}[L]}{\delta \mathbf{y}(\mathbf{x})} = \int 2(\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\mathbf{t} = 0.$$

Solving for $\mathbf{y}(\mathbf{x})$, and using the sum and product rules of probability, we obtain

$$\mathbf{y}(\mathbf{x}) = \frac{\int \mathbf{t} p(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\mathbf{t}}{\int p(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\mathbf{t}} = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) \, \mathrm{d}\mathbf{t}$$

which is the conditional average of $\mathbf{t}$ conditioned on $\mathbf{x}$. For the case of a scalar target variable we have

$$y(\mathbf{x}) = \int t p(t|\mathbf{x}) \, \mathrm{d}t$$

which is equivalent to (1.89).

1.26 **NOTE**: In the 1st printing of PRML, there is an error in equation (1.90); the integrand of the second integral should be replaced by $\mathrm{var}[t|\mathbf{x}]p(\mathbf{x})$.

We start by expanding the square in (1.151), in a similar fashion to the univariate case in the equation preceding (1.90),

$$\begin{aligned} \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 &= \|\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t}\|^2 \\ &= \|\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]\|^2 + (\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}])^{\mathrm{T}}(\mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t}) \\ &\quad + (\mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t})^{\mathrm{T}}(\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]) + \|\mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t}\|^2. \end{aligned}$$

Following the treatment of the univariate case, we now substitute this into (1.151) and perform the integral over $\mathbf{t}$. Again the cross-term vanishes and we are left with

$$\mathbb{E}[L] = \int \|\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]\|^2 p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int \mathrm{var}[\mathbf{t}|\mathbf{x}] p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

from which we see directly that the function $\mathbf{y}(\mathbf{x})$ that minimizes $\mathbb{E}[L]$ is given by $\mathbb{E}[\mathbf{t}|\mathbf{x}]$.

1.27 Since we can choose $y(\mathbf{x})$ independently for each value of $\mathbf{x}$, the minimum of the expected L_q loss can be found by minimizing the integrand given by

$$\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\,\mathrm{d}t \tag{52}$$

for each value of $\mathbf{x}$. Setting the derivative of (52) with respect to $y(\mathbf{x})$ to zero gives the stationarity condition

$$\begin{aligned}&\int q|y(\mathbf{x}) - t|^{q-1}\mathrm{sign}(y(\mathbf{x}) - t)p(t|\mathbf{x})\,\mathrm{d}t \\ &= q\int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t - q\int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t = 0\end{aligned}$$

which can also be obtained directly by setting the functional derivative of (1.91) with respect to $y(\mathbf{x})$ equal to zero. It follows that $y(\mathbf{x})$ must satisfy

$$\int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t. \tag{53}$$

For the case of $q = 1$ this reduces to

$$\int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x})\,\mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x})\,\mathrm{d}t. \tag{54}$$

which says that $y(\mathbf{x})$ must be the conditional median of t.

For $q \to 0$ we note that, as a function of t, the quantity $|y(\mathbf{x}) - t|^q$ is close to 1 everywhere except in a small neighbourhood around $t = y(\mathbf{x})$ where it falls to zero. The value of (52) will therefore be close to 1, since the density $p(t)$ is normalized, but reduced slightly by the 'notch' close to $t = y(\mathbf{x})$. We obtain the biggest reduction in (52) by choosing the location of the notch to coincide with the largest value of $p(t)$, i.e. with the (conditional) mode.

1.28 From the discussion of the introduction of Section 1.6, we have

$$h(p^2) = h(p) + h(p) = 2\,h(p).$$

We then assume that for all $k \leqslant K$, $h(p^k) = k\,h(p)$. For $k = K + 1$ we have

$$h(p^{K+1}) = h(p^K p) = h(p^K) + h(p) = K\,h(p) + h(p) = (K + 1)\,h(p).$$

Moreover,

$$h(p^{n/m}) = n\,h(p^{1/m}) = \frac{n}{m}m\,h(p^{1/m}) = \frac{n}{m}\,h(p^{m/m}) = \frac{n}{m}\,h(p)$$

and so, by continuity, we have that $h(p^x) = x\,h(p)$ for any real number x.

Now consider the positive real numbers p and q and the real number x such that $p = q^x$. From the above discussion, we see that

$$\frac{h(p)}{\ln(p)} = \frac{h(q^x)}{\ln(q^x)} = \frac{x\,h(q)}{x\,\ln(q)} = \frac{h(q)}{\ln(q)}$$

and hence $h(p) \propto \ln(p)$.

1.29 The entropy of an M-state discrete variable x can be written in the form

$$H(x) = -\sum_{i=1}^{M} p(x_i)\ln p(x_i) = \sum_{i=1}^{M} p(x_i)\ln\frac{1}{p(x_i)}. \tag{55}$$

The function $\ln(x)$ is concave$\frown$ and so we can apply Jensen's inequality in the form (1.115) but with the inequality reversed, so that

$$H(x) \leqslant \ln\left(\sum_{i=1}^{M} p(x_i)\frac{1}{p(x_i)}\right) = \ln M. \tag{56}$$

1.30 **NOTE**: In PRML, there is a minus sign ('$-$') missing on the l.h.s. of (1.103).

From (1.113) we have

$$\mathrm{KL}(p\|q) = -\int p(x)\ln q(x)\,\mathrm{d}x + \int p(x)\ln p(x)\,\mathrm{d}x. \tag{57}$$

Using (1.46) and (1.48)– (1.50), we can rewrite the first integral on the r.h.s. of (57) as

$$\begin{aligned} -\int p(x)\ln q(x)\,\mathrm{d}x &= \int \mathcal{N}(x|\mu,\sigma^2)\frac{1}{2}\left(\ln(2\pi s^2) + \frac{(x-m)^2}{s^2}\right)\mathrm{d}x \\ &= \frac{1}{2}\left(\ln(2\pi s^2) + \frac{1}{s^2}\int \mathcal{N}(x|\mu,\sigma^2)(x^2 - 2xm + m^2)\,\mathrm{d}x\right) \\ &= \frac{1}{2}\left(\ln(2\pi s^2) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2}\right). \end{aligned} \tag{58}$$

The second integral on the r.h.s. of (57) we recognize from (1.103) as the negative differential entropy of a Gaussian. Thus, from (57), (58) and (1.110), we have

$$\begin{aligned} \mathrm{KL}(p\|q) &= \frac{1}{2}\left(\ln(2\pi s^2) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2} - 1 - \ln(2\pi\sigma^2)\right) \\ &= \frac{1}{2}\left(\ln\left(\frac{s^2}{\sigma^2}\right) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2} - 1\right). \end{aligned}$$

1.31 We first make use of the relation $\mathrm{I}(\mathbf{x};\mathbf{y}) = \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{y}|\mathbf{x})$ which we obtained in (1.121), and note that the mutual information satisfies $\mathrm{I}(\mathbf{x};\mathbf{y}) \geqslant 0$ since it is a form of Kullback-Leibler divergence. Finally we make use of the relation (1.112) to obtain the desired result (1.152).

To show that statistical independence is a sufficient condition for the equality to be satisfied, we substitute $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ into the definition of the entropy, giving

$$\begin{aligned}
\mathrm{H}(\mathbf{x},\mathbf{y}) &= \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{x},\mathbf{y}) \,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\
&= \iint p(\mathbf{x})p(\mathbf{y}) \left\{\ln p(\mathbf{x}) + \ln p(\mathbf{y})\right\} \,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{x}) \ln p(\mathbf{x}) \,\mathrm{d}\mathbf{x} + \int p(\mathbf{y}) \ln p(\mathbf{y}) \,\mathrm{d}\mathbf{y} \\
&= \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}).
\end{aligned}$$

To show that statistical independence is a necessary condition, we combine the equality condition

$$\mathrm{H}(\mathbf{x},\mathbf{y}) = \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y})$$

with the result (1.112) to give

$$\mathrm{H}(\mathbf{y}|\mathbf{x}) = \mathrm{H}(\mathbf{y}).$$

We now note that the right-hand side is independent of $\mathbf{x}$ and hence the left-hand side must also be constant with respect to $\mathbf{x}$. Using (1.121) it then follows that the mutual information $\mathrm{I}[\mathbf{x},\mathbf{y}] = 0$. Finally, using (1.120) we see that the mutual information is a form of KL divergence, and this vanishes only if the two distributions are equal, so that $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ as required.

1.32 When we make a change of variables, the probability density is transformed by the Jacobian of the change of variables. Thus we have

$$p(\mathbf{x}) = p(\mathbf{y}) \left|\frac{\partial y_i}{\partial x_j}\right| = p(\mathbf{y})|\mathbf{A}| \tag{59}$$

where $|\cdot|$ denotes the determinant. Then the entropy of $\mathbf{y}$ can be written

$$H(\mathbf{y}) = -\int p(\mathbf{y}) \ln p(\mathbf{y}) \,\mathrm{d}\mathbf{y} = -\int p(\mathbf{x}) \ln \left\{p(\mathbf{x})|\mathbf{A}|^{-1}\right\} \,\mathrm{d}\mathbf{x} = H(\mathbf{x}) + \ln|\mathbf{A}| \tag{60}$$

as required.

1.33 The conditional entropy $H(y|x)$ can be written

$$H(y|x) = -\sum_i \sum_j p(y_i|x_j)p(x_j) \ln p(y_i|x_j) \tag{61}$$

which equals 0 by assumption. Since the quantity $-p(y_i|x_j)\ln p(y_i|x_j)$ is non-negative each of these terms must vanish for any value x_j such that $p(x_j) \neq 0$. However, the quantity $p \ln p$ only vanishes for $p = 0$ or $p = 1$. Thus the quantities $p(y_i|x_j)$ are all either 0 or 1. However, they must also sum to 1, since this is a normalized probability distribution, and so precisely one of the $p(y_i|x_j)$ is 1, and the rest are 0. Thus, for each value x_j there is a unique value y_i with non-zero probability.

1.34 Obtaining the required functional derivative can be done simply by inspection. However, if a more formal approach is required we can proceed as follows using the techniques set out in Appendix D. Consider first the functional

$$I[p(x)] = \int p(x)f(x)\,\mathrm{d}x.$$

Under a small variation $p(x) \to p(x) + \epsilon\eta(x)$ we have

$$I[p(x) + \epsilon\eta(x)] = \int p(x)f(x)\,\mathrm{d}x + \epsilon \int \eta(x)f(x)\,\mathrm{d}x$$

and hence from (D.3) we deduce that the functional derivative is given by

$$\frac{\delta I}{\delta p(x)} = f(x).$$

Similarly, if we define

$$J[p(x)] = \int p(x)\ln p(x)\,\mathrm{d}x$$

then under a small variation $p(x) \to p(x) + \epsilon\eta(x)$ we have

$$\begin{aligned} J[p(x) + \epsilon\eta(x)] &= \int p(x)\ln p(x)\,\mathrm{d}x \\ &\quad + \epsilon\left\{\int \eta(x)\ln p(x)\,\mathrm{d}x + \int p(x)\frac{1}{p(x)}\eta(x)\,\mathrm{d}x\right\} + O(\epsilon^2) \end{aligned}$$

and hence

$$\frac{\delta J}{\delta p(x)} = p(x) + 1.$$

Using these two results we obtain the following result for the functional derivative

$$-\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2.$$

Re-arranging then gives (1.108).

To eliminate the Lagrange multipliers we substitute (1.108) into each of the three constraints (1.105), (1.106) and (1.107) in turn. The solution is most easily obtained

by comparison with the standard form of the Gaussian, and noting that the results

$$\lambda_1 = 1 - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \tag{62}$$

$$\lambda_2 = 0 \tag{63}$$

$$\lambda_3 = \frac{1}{2\sigma^2} \tag{64}$$

do indeed satisfy the three constraints.

Note that there is a typographical error in the question, which should read "Use calculus of variations to show that the stationary point of the functional shown just before (1.108) is given by (1.108)".

For the multivariate version of this derivation, see Exercise 2.14.

1.35 **NOTE**: In PRML, there is a minus sign ('$-$') missing on the l.h.s. of (1.103).

Substituting the right hand side of (1.109) in the argument of the logarithm on the right hand side of (1.103), we obtain

$$\begin{aligned}
\mathrm{H}[x] &= -\int p(x)\ln p(x)\,\mathrm{d}x \\
&= -\int p(x)\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\
&= \frac{1}{2}\left(\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\int p(x)(x-\mu)^2\,\mathrm{d}x\right) \\
&= \frac{1}{2}\left(\ln(2\pi\sigma^2) + 1\right),
\end{aligned}$$

where in the last step we used (1.107).

1.36 Consider (1.114) with $\lambda = 0.5$ and $b = a + 2\epsilon$ (and hence $a = b - 2\epsilon$),

$$\begin{aligned}
0.5f(a) + 0.5f(b) &> f(0.5a + 0.5b) \\
&= 0.5f(0.5a + 0.5(a + 2\epsilon)) + 0.5f(0.5(b - 2\epsilon) + 0.5b) \\
&= 0.5f(a + \epsilon) + 0.5f(b - \epsilon)
\end{aligned}$$

We can rewrite this as

$$f(b) - f(b - \epsilon) > f(a + \epsilon) - f(a)$$

We then divide both sides by ϵ and let $\epsilon \to 0$, giving

$$f'(b) > f'(a).$$

Since this holds at all points, it follows that $f''(x) \geqslant 0$ everywhere.

To show the implication in the other direction, we make use of Taylor's theorem (with the remainder in Lagrange form), according to which there exist an $x^\star$ such that

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x^\star)(x - x_0)^2.$$

Since we assume that $f''(x) > 0$ everywhere, the third term on the r.h.s. will always be positive and therefore

$$f(x) > f(x_0) + f'(x_0)(x - x_0)$$

Now let $x_0 = \lambda a + (1 - \lambda)b$ and consider setting $x = a$, which gives

$$\begin{aligned} f(a) &> f(x_0) + f'(x_0)(a - x_0) \\ &= f(x_0) + f'(x_0)\left((1 - \lambda)(a - b)\right). \end{aligned} \tag{65}$$

Similarly, setting $x = b$ gives

$$f(b) > f(x_0) + f'(x_0)(\lambda(b - a)). \tag{66}$$

Multiplying (65) by λ and (66) by $1 - \lambda$ and adding up the results on both sides, we obtain

$$\lambda f(a) + (1 - \lambda)f(b) > f(x_0) = f(\lambda a + (1 - \lambda)b)$$

as required.

1.37 From (1.104), making use of (1.111), we have

$$\begin{aligned} \mathrm{H}[\mathbf{x}, \mathbf{y}] &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\right) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})\right) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} - \int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]. \end{aligned}$$

1.38 From (1.114) we know that the result (1.115) holds for $M = 1$. We now suppose that it holds for some general value M and show that it must therefore hold for $M + 1$. Consider the left hand side of (1.115)

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) = f\left(\lambda_{M+1} x_{M+1} + \sum_{i=1}^{M} \lambda_i x_i\right) \tag{67}$$

$$= f\left(\lambda_{M+1} x_{M+1} + (1 - \lambda_{M+1}) \sum_{i=1}^{M} \eta_i x_i\right) \tag{68}$$

where we have defined

$$\eta_i = \frac{\lambda_i}{1-\lambda_{M+1}}. \tag{69}$$

We now apply (1.114) to give

$$f\left(\sum_{i=1}^{M+1}\lambda_i x_i\right) \leqslant \lambda_{M+1} f(x_{M+1}) + (1-\lambda_{M+1}) f\left(\sum_{i=1}^{M}\eta_i x_i\right). \tag{70}$$

We now note that the quantities λ_i by definition satisfy

$$\sum_{i=1}^{M+1}\lambda_i = 1 \tag{71}$$

and hence we have

$$\sum_{i=1}^{M}\lambda_i = 1-\lambda_{M+1} \tag{72}$$

Then using (69) we see that the quantities η_i satisfy the property

$$\sum_{i=1}^{M}\eta_i = \frac{1}{1-\lambda_{M+1}}\sum_{i=1}^{M}\lambda_i = 1. \tag{73}$$

Thus we can apply the result (1.115) at order M and so (70) becomes

$$f\left(\sum_{i=1}^{M+1}\lambda_i x_i\right) \leqslant \lambda_{M+1} f(x_{M+1}) + (1-\lambda_{M+1})\sum_{i=1}^{M}\eta_i f(x_i) = \sum_{i=1}^{M+1}\lambda_i f(x_i) \tag{74}$$

where we have made use of (69).

1.39 From Table 1.3 we obtain the marginal probabilities by summation and the conditional probabilities by normalization, to give

x	
0	2/3
1	1/3

$p(x)$

y	
0	1
1/3	2/3

$p(y)$

x \ y	0	1
0	1	1/2
1	0	1/2

$p(x|y)$

x \ y	0	1
0	1/2	1/2
1	0	1

$p(y|x)$

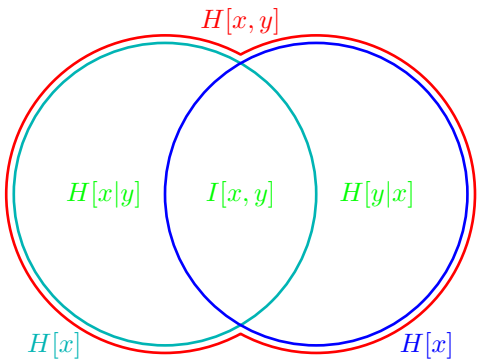


Figure 2 Diagram showing the relationship between marginal, conditional and joint entropies and the mutual information.

From these tables, together with the definitions

$$H(x) = -\sum_i p(x_i) \ln p(x_i) \tag{75}$$

$$H(x|y) = -\sum_i \sum_j p(x_i, y_j) \ln p(x_i|y_j) \tag{76}$$

and similar definitions for $H(y)$ and $H(y|x)$, we obtain the following results

(a) $H(x) = \ln 3 - \frac{2}{3} \ln 2$

(b) $H(y) = \ln 3 - \frac{2}{3} \ln 2$

(c) $H(y|x) = \frac{2}{3} \ln 2$

(d) $H(x|y) = \frac{2}{3} \ln 2$

(e) $H(x, y) = \ln 3$

(f) $I(x; y) = \ln 3 - \frac{4}{3} \ln 2$

where we have used (1.121) to evaluate the mutual information. The corresponding diagram is shown in Figure 2.

1.40 The arithmetic and geometric means are defined as

$$\bar{x}_{\mathrm{A}} = \frac{1}{K} \sum_k^K x_k \quad \text{and} \quad \bar{x}_{\mathrm{G}} = \left(\prod_k^K x_k \right)^{1/K},$$

respectively. Taking the logarithm of $\bar{x}_{\mathrm{A}}$ and $\bar{x}_{\mathrm{G}}$, we see that

$$\ln \bar{x}_{\mathrm{A}} = \ln \left(\frac{1}{K} \sum_k^K x_k \right) \quad \text{and} \quad \ln \bar{x}_{\mathrm{G}} = \frac{1}{K} \sum_k^K \ln x_k.$$

By matching f with $\ln$ and λ_i with $1/K$ in (1.115), taking into account that the logarithm is concave rather than convex and the inequality therefore goes the other way, we obtain the desired result.

1.41 From the product rule we have $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$, and so (1.120) can be written as

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &= -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\int p(\mathbf{y}) \ln p(\mathbf{y})\, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \end{aligned} \tag{77}$$

Chapter 2 Probability Distributions

2.1 From the definition (2.2) of the Bernoulli distribution we have

$$\begin{aligned} \sum_{x\in\{0,1\}} p(x|\mu) &= p(x=0|\mu) + p(x=1|\mu) \\ &= (1-\mu) + \mu = 1 \\ \sum_{x\in\{0,1\}} x p(x|\mu) &= 0.p(x=0|\mu) + 1.p(x=1|\mu) = \mu \\ \sum_{x\in\{0,1\}} (x-\mu)^2 p(x|\mu) &= \mu^2 p(x=0|\mu) + (1-\mu)^2 p(x=1|\mu) \\ &= \mu^2(1-\mu) + (1-\mu)^2\mu = \mu(1-\mu). \end{aligned}$$

The entropy is given by

$$\begin{aligned} \mathrm{H}[x] &= -\sum_{x\in\{0,1\}} p(x|\mu) \ln p(x|\mu) \\ &= -\sum_{x\in\{0,1\}} \mu^x (1-\mu)^{1-x} \{x\ln\mu + (1-x)\ln(1-\mu)\} \\ &= -(1-\mu)\ln(1-\mu) - \mu\ln\mu. \end{aligned}$$

2.2 The normalization of (2.261) follows from

$$p(x=+1|\mu) + p(x=-1|\mu) = \left(\frac{1+\mu}{2}\right) + \left(\frac{1-\mu}{2}\right) = 1.$$

The mean is given by

$$\mathbb{E}[x] = \left(\frac{1+\mu}{2}\right) - \left(\frac{1-\mu}{2}\right) = \mu.$$

To evaluate the variance we use

$$\mathbb{E}[x^2] = \left(\frac{1-\mu}{2}\right) + \left(\frac{1+\mu}{2}\right) = 1$$

from which we have

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = 1 - \mu^2.$$

Finally the entropy is given by

$$\begin{aligned} \mathrm{H}[x] &= -\sum_{x=-1}^{x=+1} p(x|\mu) \ln p(x|\mu) \\ &= -\left(\frac{1-\mu}{2}\right) \ln \left(\frac{1-\mu}{2}\right) - \left(\frac{1+\mu}{2}\right) \ln \left(\frac{1+\mu}{2}\right). \end{aligned}$$

2.3 Using the definition (2.10) we have

$$\begin{aligned} \binom{N}{n} + \binom{N}{n-1} &= \frac{N!}{n!(N-n)!} + \frac{N!}{(n-1)!(N+1-n)!} \\ &= \frac{(N+1-n)N! + nN!}{n!(N+1-n)!} = \frac{(N+1)!}{n!(N+1-n)!} \\ &= \binom{N+1}{n}. \end{aligned} \tag{78}$$

To prove the binomial theorem (2.263) we note that the theorem is trivially true for $N = 0$. We now assume that it holds for some general value N and prove its correctness for $N + 1$, which can be done as follows

$$\begin{aligned} (1+x)^{N+1} &= (1+x) \sum_{n=0}^{N} \binom{N}{n} x^n \\ &= \sum_{n=0}^{N} \binom{N}{n} x^n + \sum_{n=1}^{N+1} \binom{N}{n-1} x^n \\ &= \binom{N}{0} x^0 + \sum_{n=1}^{N} \left\{ \binom{N}{n} + \binom{N}{n-1} \right\} x^n + \binom{N}{N} x^{N+1} \\ &= \binom{N+1}{0} x^0 + \sum_{n=1}^{N} \binom{N+1}{n} x^n + \binom{N+1}{N+1} x^{N+1} \\ &= \sum_{n=0}^{N+1} \binom{N+1}{n} x^n \end{aligned} \tag{79}$$

which completes the inductive proof. Finally, using the binomial theorem, the normalization condition (2.264) for the binomial distribution gives

$$\begin{aligned}\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n} &= (1-\mu)^N\sum_{n=0}^{N}\binom{N}{n}\left(\frac{\mu}{1-\mu}\right)^n \\ &= (1-\mu)^N\left(1+\frac{\mu}{1-\mu}\right)^N = 1 \qquad (80)\end{aligned}$$

as required.

2.4 Differentiating (2.264) with respect to μ we obtain

$$\sum_{n=1}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\frac{n}{\mu}-\frac{(N-n)}{(1-\mu)}\right]=0.$$

Multiplying through by $\mu(1-\mu)$ and re-arranging we obtain (2.11).

If we differentiate (2.264) twice with respect to μ we obtain

$$\sum_{n=1}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left\{\left[\frac{n}{\mu}-\frac{(N-n)}{(1-\mu)}\right]^2-\frac{n}{\mu^2}-\frac{(N-n)}{(1-\mu)^2}\right\}=0.$$

We now multiply through by $\mu^2(1-\mu)^2$ and re-arrange, making use of the result (2.11) for the mean of the binomial distribution, to obtain

$$\mathbb{E}[n^2]=N\mu(1-\mu)+N^2\mu^2.$$

Finally, we use (1.40) to obtain the result (2.12) for the variance.

2.5 Making the change of variable $t=y+x$ in (2.266) we obtain

$$\Gamma(a)\Gamma(b)=\int_0^\infty x^{a-1}\left\{\int_x^\infty \exp(-t)(t-x)^{b-1}\,\mathrm{d}t\right\}\mathrm{d}x. \qquad (81)$$

We now exchange the order of integration, taking care over the limits of integration

$$\Gamma(a)\Gamma(b)=\int_0^\infty\int_0^t x^{a-1}\exp(-t)(t-x)^{b-1}\,\mathrm{d}x\,\mathrm{d}t. \qquad (82)$$

The change in the limits of integration in going from (81) to (82) can be understood by reference to Figure 3. Finally we change variables in the x integral using $x=t\mu$ to give

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty \exp(-t)t^{a-1}t^{b-1}t\,\mathrm{d}t\int_0^1\mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu \\ &= \Gamma(a+b)\int_0^1\mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu. \qquad (83)\end{aligned}$$

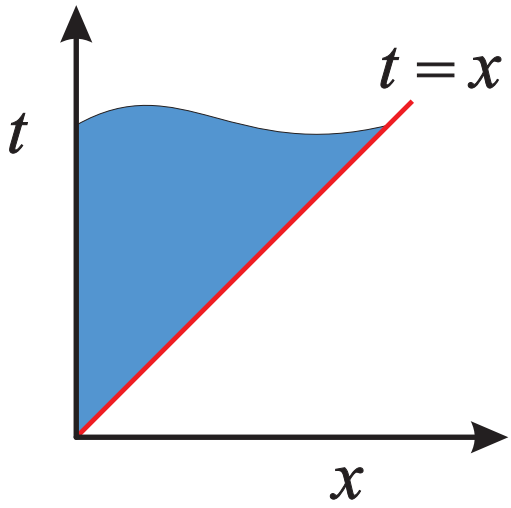


Figure 3 Plot of the region of integration of (81) in (x, t) space.

2.6 From (2.13) the mean of the beta distribution is given by

$$\mathbb{E}[\mu] = \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{(a+1)-1}(1-\mu)^{b-1}\,\mathrm{d}\mu.$$

Using the result (2.265), which follows directly from the normalization condition for the Beta distribution, we have

$$\mathbb{E}[\mu] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)} = \frac{a}{a+b}$$

where we have used the property $\Gamma(x+1) = x\Gamma(x)$. We can find the variance in the same way, by first showing that

$$\begin{aligned} \mathbb{E}[\mu^2] &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \frac{\Gamma(a+2+b)}{\Gamma(a+2)\Gamma(b)} \mu^{(a+2)-1}(1-\mu)^{b-1}\,\mathrm{d}\mu \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2+b)}{\Gamma(a+2)\Gamma(b)} = \frac{a}{(a+b)} \frac{a+1}{(a+1+b)}. \end{aligned} \tag{84}$$

Now we use the result (1.40), together with the result (2.15) to derive the result (2.16) for $\mathrm{var}[\mu]$. Finally, we obtain the result (2.269) for the mode of the beta distribution simply by setting the derivative of the right hand side of (2.13) with respect to μ to zero and re-arranging.

2.7 **NOTE**: In PRML, the exercise text contains a typographical error. On the third line, "mean value of x" should be "mean value of μ".

Using the result (2.15) for the mean of a Beta distribution we see that the prior mean is $a/(a+b)$ while the posterior mean is $(a+n)/(a+b+n+m)$. The maximum likelihood estimate for μ is given by the relative frequency $n/(n+m)$ of observations

of $x = 1$. Thus the posterior mean will lie between the prior mean and the maximum likelihood solution provided the following equation is satisfied for λ in the interval $(0, 1)$

$$\lambda \frac{a}{a+b} + (1-\lambda)\frac{n}{n+m} = \frac{a+n}{a+b+n+m}.$$

which represents a convex combination of the prior mean and the maximum likelihood estimator. This is a linear equation for λ which is easily solved by re-arranging terms to give

$$\lambda = \frac{1}{1 + (n+m)/(a+b)}.$$

Since $a > 0$, $b > 0$, $n > 0$, and $m > 0$, it follows that the term $(n+m)/(a+b)$ lies in the range $(0, \infty)$ and hence λ must lie in the range $(0, 1)$.

2.8 To prove the result (2.270) we use the product rule of probability

$$\begin{aligned} \mathbb{E}_y\left[\mathbb{E}_x[x|y]\right] &= \int \left\{ \int x p(x|y)\,\mathrm{d}x \right\} p(y)\,\mathrm{d}y \\ &= \iint x p(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int x p(x)\,\mathrm{d}x = \mathbb{E}_x[x]. \end{aligned} \tag{85}$$

For the result (2.271) for the conditional variance we make use of the result (1.40), as well as the relation (85), to give

$$\begin{aligned} \mathbb{E}_y\left[\mathrm{var}_x[x|y]\right] + \mathrm{var}_y\left[\mathbb{E}_x[x|y]\right] &= \mathbb{E}_y\left[\mathbb{E}_x[x^2|y] - \mathbb{E}_x[x|y]^2\right] \\ &\quad + \mathbb{E}_y\left[\mathbb{E}_x[x|y]^2\right] - \mathbb{E}_y\left[\mathbb{E}_x[x|y]\right]^2 \\ &= \mathbb{E}_x[x^2] - \mathbb{E}_x[x]^2 = \mathrm{var}_x[x] \end{aligned}$$

where we have made use of $\mathbb{E}_y\left[\mathbb{E}_x[x^2|y]\right] = \mathbb{E}_x[x^2]$ which can be proved by analogy with (85).

2.9 When we integrate over μ_{M-1} the lower limit of integration is 0, while the upper limit is $1 - \sum_{j=1}^{M-2} \mu_j$ since the remaining probabilities must sum to one (see Figure 2.4). Thus we have

$$\begin{aligned} p_{M-1}(\mu_1, \ldots, \mu_{M-2}) &= \int_0^{1-\sum_{j=1}^{M-2}\mu_j} p_M(\mu_1, \ldots, \mu_{M-1})\,\mathrm{d}\mu_{M-1} \\ &= C_M \left[\prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \right] \int_0^{1-\sum_{j=1}^{M-2}\mu_j} \mu_{M-1}^{\alpha_{M-1}-1} \left(1 - \sum_{j=1}^{M-1} \mu_j \right)^{\alpha_M - 1} \mathrm{d}\mu_{M-1}. \end{aligned}$$

In order to make the limits of integration equal to 0 and 1 we change integration variable from μ_{M-1} to t using

$$\mu_{M-1} = t\left(1 - \sum_{j=1}^{M-2} \mu_j\right)$$

which gives

$$\begin{aligned} & p_{M-1}(\mu_1,\ldots,\mu_{M-2}) \\ &= C_M\left[\prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\right]\left(1-\sum_{j=1}^{M-2}\mu_j\right)^{\alpha_{M-1}+\alpha_M-1}\int_0^1 t^{\alpha_{M-1}-1}(1-t)^{\alpha_M-1}\,\mathrm{d}t \\ &= C_M\left[\prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\right]\left(1-\sum_{j=1}^{M-2}\mu_j\right)^{\alpha_{M-1}+\alpha_M-1}\frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)} \end{aligned} \tag{86}$$

where we have used (2.265). The right hand side of (86) is seen to be a normalized Dirichlet distribution over $M-1$ variables, with coefficients $\alpha_1,\ldots,\alpha_{M-2},\alpha_{M-1}+\alpha_M$, (note that we have effectively combined the final two categories) and we can identify its normalization coefficient using (2.38). Thus

$$\begin{aligned} C_M &= \frac{\Gamma(\alpha_1+\ldots+\alpha_M)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_{M-2})\Gamma(\alpha_{M-1}+\alpha_M)}\cdot\frac{\Gamma(\alpha_{M-1}+\alpha_M)}{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)} \\ &= \frac{\Gamma(\alpha_1+\ldots+\alpha_M)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_M)} \end{aligned} \tag{87}$$

as required.

2.10 Using the fact that the Dirichlet distribution (2.38) is normalized we have

$$\int\prod_{k=1}^{M}\mu_k^{\alpha_k-1}\,\mathrm{d}\boldsymbol{\mu}=\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_M)}{\Gamma(\alpha_0)} \tag{88}$$

where $\int \mathrm{d}\boldsymbol{\mu}$ denotes the integral over the $(M-1)$-dimensional simplex defined by $0\leqslant\mu_k\leqslant 1$ and $\sum_k\mu_k=1$. Now consider the expectation of μ_j which can be written

$$\begin{aligned} \mathbb{E}[\mu_j] &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_M)}\int\mu_j\prod_{k=1}^{M}\mu_k^{\alpha_k-1}\,\mathrm{d}\boldsymbol{\mu} \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_M)}\cdot\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_j+1)\cdots\Gamma(\alpha_M)}{\Gamma(\alpha_0+1)}=\frac{\alpha_j}{\alpha_0} \end{aligned}$$

where we have made use of (88), noting that the effect of the extra factor of μ_j is to increase the coefficient α_j by 1, and then made use of $\Gamma(x+1)=x\Gamma(x)$. By similar reasoning we have

$$\begin{aligned} \mathrm{var}[\mu_j] &= \mathbb{E}[\mu_j^2]-\mathbb{E}[\mu_j]^2=\frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)}-\frac{\alpha_j^2}{\alpha_0^2} \\ &= \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}. \end{aligned}$$

Likewise, for $j \neq l$ we have

$$\begin{aligned}\mathrm{cov}[\mu_j \mu_l] &= \mathbb{E}[\mu_j \mu_l] - \mathbb{E}[\mu_j]\mathbb{E}[\mu_l] = \frac{\alpha_j \alpha_l}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j}{\alpha_0}\frac{\alpha_l}{\alpha_0} \\ &= -\frac{\alpha_j \alpha_l}{\alpha_0^2(\alpha_0+1)}.\end{aligned}$$

2.11 We first of all write the Dirichlet distribution (2.38) in the form

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = K(\boldsymbol{\alpha}) \prod_{k=1}^{M} \mu_k^{\alpha_k - 1}$$

where

$$K(\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_M)}.$$

Next we note the following relation

$$\begin{aligned}\frac{\partial}{\partial \alpha_j}\prod_{k=1}^{M} \mu_k^{\alpha_k - 1} &= \frac{\partial}{\partial \alpha_j}\prod_{k=1}^{M} \exp\left((\alpha_k - 1)\ln \mu_k\right) \\ &= \prod_{k=1}^{M} \ln \mu_j \exp\left\{(\alpha_k - 1)\ln \mu_k\right\} \\ &= \ln \mu_j \prod_{k=1}^{M} \mu_k^{\alpha_k - 1}\end{aligned}$$

from which we obtain

$$\begin{aligned}\mathbb{E}[\ln \mu_j] &= K(\boldsymbol{\alpha}) \int_0^1 \cdots \int_0^1 \ln \mu_j \prod_{k=1}^{M} \mu_k^{\alpha_k - 1}\, \mathrm{d}\mu_1 \ldots \mathrm{d}\mu_M \\ &= K(\boldsymbol{\alpha}) \frac{\partial}{\partial \alpha_j} \int_0^1 \cdots \int_0^1 \prod_{k=1}^{M} \mu_k^{\alpha_k - 1}\, \mathrm{d}\mu_1 \ldots \mathrm{d}\mu_M \\ &= K(\boldsymbol{\alpha}) \frac{\partial}{\partial \mu_j} \frac{1}{K(\boldsymbol{\alpha})} \\ &= -\frac{\partial}{\partial \mu_j} \ln K(\boldsymbol{\alpha}).\end{aligned}$$

Finally, using the expression for $K(\boldsymbol{\alpha})$, together with the definition of the digamma function $\psi(\cdot)$, we have

$$\mathbb{E}[\ln \mu_j] = \psi(\alpha_j) - \psi(\alpha_0).$$

2.12 The normalization of the uniform distribution is proved trivially

$$\int_a^b \frac{1}{b-a}\,\mathrm{d}x = \frac{b-a}{b-a} = 1.$$

For the mean of the distribution we have

$$\mathbb{E}[x] = \int_a^b \frac{1}{b-a}x\,\mathrm{d}x = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

The variance can be found by first evaluating

$$\mathbb{E}[x^2] = \int_a^b \frac{1}{b-a}x^2\,\mathrm{d}x = \left[\frac{x^3}{3(b-a)}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}$$

and then using (1.40) to give

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

2.13 Note that this solution is the multivariate version of Solution 1.30.

From (1.113) we have

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Using (2.43), (2.57), (2.59) and (2.62), we can rewrite the first integral on the r.h.s. of () as

$$\begin{aligned}
&-\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&\quad= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}^2)\frac{1}{2}\left(D\ln(2\pi) + \ln|\mathbf{L}| + (\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})\right)\,\mathrm{d}\mathbf{x} \\
&\quad= \frac{1}{2}\left(D\ln(2\pi) + \ln|\mathbf{L}| + \mathrm{Tr}[\mathbf{L}^{-1}(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma})]\right. \\
&\qquad\qquad \left. -\boldsymbol{\mu}\mathbf{L}^{-1}\mathbf{m} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu} + \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m}\right). \qquad (89)
\end{aligned}$$

The second integral on the r.h.s. of () we recognize from (1.104) as the negative differential entropy of a multivariate Gaussian. Thus, from (), (89) and (B.41), we have

$$\begin{aligned}
\mathrm{KL}(p\|q) = \frac{1}{2}\Bigg(&\ln\frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} + \mathrm{Tr}[\mathbf{L}^{-1}(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma})] \\
&- \boldsymbol{\mu}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu} + \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} - D\Bigg)
\end{aligned}$$

2.14 As for the univariate Gaussian considered in Section 1.6, we can make use of Lagrange multipliers to enforce the constraints on the maximum entropy solution. Note that we need a single Lagrange multiplier for the normalization constraint (2.280), a D-dimensional vector $\mathbf{m}$ of Lagrange multipliers for the D constraints given by (2.281), and a $D \times D$ matrix $\mathbf{L}$ of Lagrange multipliers to enforce the D^2 constraints represented by (2.282). Thus we maximize

$$\begin{aligned}\widetilde{\mathrm{H}}[p] = & -\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \lambda \left(\int p(\mathbf{x}) \, \mathrm{d}\mathbf{x} - 1 \right) \\ & + \mathbf{m}^{\mathrm{T}} \left(\int p(\mathbf{x}) \mathbf{x} \, \mathrm{d}\mathbf{x} - \boldsymbol{\mu} \right) \\ & + \mathrm{Tr} \left\{ \mathbf{L} \left(\int p(\mathbf{x}) (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \, \mathrm{d}\mathbf{x} - \boldsymbol{\Sigma} \right) \right\}. \end{aligned} \tag{90}$$

By functional differentiation (Appendix D) the maximum of this functional with respect to $p(\mathbf{x})$ occurs when

$$0 = -1 - \ln p(\mathbf{x}) + \lambda + \mathbf{m}^{\mathrm{T}} \mathbf{x} + \mathrm{Tr}\{\mathbf{L}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\}.$$

Solving for $p(\mathbf{x})$ we obtain

$$p(\mathbf{x}) = \exp \left\{ \lambda - 1 + \mathbf{m}^{\mathrm{T}} \mathbf{x} + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{L} (\mathbf{x} - \boldsymbol{\mu}) \right\}. \tag{91}$$

We now find the values of the Lagrange multipliers by applying the constraints. First we complete the square inside the exponential, which becomes

$$\lambda - 1 + \left(\mathbf{x} - \boldsymbol{\mu} + \frac{1}{2} \mathbf{L}^{-1} \mathbf{m} \right)^{\mathrm{T}} \mathbf{L} \left(\mathbf{x} - \boldsymbol{\mu} + \frac{1}{2} \mathbf{L}^{-1} \mathbf{m} \right) + \boldsymbol{\mu}^{\mathrm{T}} \mathbf{m} - \frac{1}{4} \mathbf{m}^{\mathrm{T}} \mathbf{L}^{-1} \mathbf{m}.$$

We now make the change of variable

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} + \frac{1}{2} \mathbf{L}^{-1} \mathbf{m}.$$

The constraint (2.281) then becomes

$$\int \exp \left\{ \lambda - 1 + \mathbf{y}^{\mathrm{T}} \mathbf{L} \mathbf{y} + \boldsymbol{\mu}^{\mathrm{T}} \mathbf{m} - \frac{1}{4} \mathbf{m}^{\mathrm{T}} \mathbf{L}^{-1} \mathbf{m} \right\} \left(\mathbf{y} + \boldsymbol{\mu} - \frac{1}{2} \mathbf{L}^{-1} \mathbf{m} \right) \mathrm{d}\mathbf{y} = \boldsymbol{\mu}.$$

In the final parentheses, the term in $\mathbf{y}$ vanishes by symmetry, while the term in $\boldsymbol{\mu}$ simply integrates to $\boldsymbol{\mu}$ by virtue of the normalization constraint (2.280) which now takes the form

$$\int \exp \left\{ \lambda - 1 + \mathbf{y}^{\mathrm{T}} \mathbf{L} \mathbf{y} + \boldsymbol{\mu}^{\mathrm{T}} \mathbf{m} - \frac{1}{4} \mathbf{m}^{\mathrm{T}} \mathbf{L}^{-1} \mathbf{m} \right\} \mathrm{d}\mathbf{y} = 1.$$

and hence we have

$$-\frac{1}{2} \mathbf{L}^{-1} \mathbf{m} = \mathbf{0}$$

where again we have made use of the constraint (2.280). Thus $\mathbf{m} = \mathbf{0}$ and so the density becomes

$$p(\mathbf{x}) = \exp\left\{\lambda - 1 + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Substituting this into the final constraint (2.282), and making the change of variable $\mathbf{x} - \boldsymbol{\mu} = \mathbf{z}$ we obtain

$$\int \exp\left\{\lambda - 1 + \mathbf{z}^{\mathrm{T}}\mathbf{L}\mathbf{z}\right\}\mathbf{z}\mathbf{z}^{\mathrm{T}}\,\mathrm{d}\mathbf{x} = \boldsymbol{\Sigma}.$$

Applying an analogous argument to that used to derive (2.64) we obtain $\mathbf{L} = -\frac{1}{2}\boldsymbol{\Sigma}$. Finally, the value of λ is simply that value needed to ensure that the Gaussian distribution is correctly normalized, as derived in Section 2.3, and hence is given by

$$\lambda - 1 = \ln\left\{\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\right\}.$$

2.15 From the definitions of the multivariate differential entropy (1.104) and the multivariate Gaussian distribution (2.43), we get

$$\begin{aligned}
\mathrm{H}[\mathbf{x}] &= -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\ln\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\,\mathrm{d}\mathbf{x} \\
&= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{1}{2}\left(D\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\,\mathrm{d}\mathbf{x} \\
&= \frac{1}{2}\left(D\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right]\right) \\
&= \frac{1}{2}\left(D\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + D\right)
\end{aligned}$$

2.16 We have $p(x_1) = \mathcal{N}(x_1|\mu_1,\tau_1^{-1})$ and $p(x_2) = \mathcal{N}(x_2|\mu_2,\tau_2^{-1})$. Since $x = x_1 + x_2$ we also have $p(x|x_2) = \mathcal{N}(x|\mu_1 + x_2, \tau_1^{-1})$. We now evaluate the convolution integral given by (2.284) which takes the form

$$p(x) = \left(\frac{\tau_1}{2\pi}\right)^{1/2}\left(\frac{\tau_2}{2\pi}\right)^{1/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{\tau_1}{2}(x-\mu_1-x_2)^2 - \frac{\tau_2}{2}(x_2-\mu_2)^2\right\}\,\mathrm{d}x_2. \tag{92}$$

Since the final result will be a Gaussian distribution for $p(x)$ we need only evaluate its precision, since, from (1.110), the entropy is determined by the variance or equivalently the precision, and is independent of the mean. This allows us to simplify the calculation by ignoring such things as normalization constants.

We begin by considering the terms in the exponent of (92) which depend on x_2 which are given by

$$\begin{aligned}
&-\frac{1}{2}x_2^2(\tau_1+\tau_2) + x_2\left\{\tau_1(x-\mu_1) + \tau_2\mu_2\right\} \\
&\quad = -\frac{1}{2}(\tau_1+\tau_2)\left\{x_2 - \frac{\tau_1(x-\mu_1)+\tau_2\mu_2}{\tau_1+\tau_2}\right\}^2 + \frac{\left\{\tau_1(x-\mu_1)+\tau_2\mu_2\right\}^2}{2(\tau_1+\tau_2)}
\end{aligned}$$

where we have completed the square over x_2. When we integrate out x_2, the first term on the right hand side will simply give rise to a constant factor independent of x. The second term, when expanded out, will involve a term in x^2. Since the precision of x is given directly in terms of the coefficient of x^2 in the exponent, it is only such terms that we need to consider. There is one other term in x^2 arising from the original exponent in (92). Combining these we have

$$-\frac{\tau_1}{2}x^2 + \frac{\tau_1^2}{2(\tau_1+\tau_2)}x^2 = -\frac{1}{2}\frac{\tau_1\tau_2}{\tau_1+\tau_2}x^2$$

from which we see that x has precision $\tau_1\tau_2/(\tau_1+\tau_2)$.

We can also obtain this result for the precision directly by appealing to the general result (2.115) for the convolution of two linear-Gaussian distributions.

The entropy of x is then given, from (1.110), by

$$\mathrm{H}[x] = \frac{1}{2}\ln\left\{\frac{2\pi(\tau_1+\tau_2)}{\tau_1\tau_2}\right\}.$$

2.17 We can use an analogous argument to that used in the solution of Exercise 1.14. Consider a general square matrix $\mathbf{\Lambda}$ with elements Λ_{ij}. Then we can always write $\mathbf{\Lambda} = \mathbf{\Lambda}^{\mathrm{A}} + \mathbf{\Lambda}^{\mathrm{S}}$ where

$$\Lambda_{ij}^{\mathrm{S}} = \frac{\Lambda_{ij}+\Lambda_{ji}}{2}, \qquad \Lambda_{ij}^{\mathrm{A}} = \frac{\Lambda_{ij}-\Lambda_{ji}}{2} \tag{93}$$

and it is easily verified that $\mathbf{\Lambda}^{\mathrm{S}}$ is symmetric so that $\Lambda_{ij}^{\mathrm{S}} = \Lambda_{ji}^{\mathrm{S}}$, and $\mathbf{\Lambda}^{\mathrm{A}}$ is antisymmetric so that $\Lambda_{ij}^{\mathrm{A}} = -\Lambda_{ji}^{\mathrm{S}}$. The quadratic form in the exponent of a D-dimensional multivariate Gaussian distribution can be written

$$\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(x_i-\mu_i)\Lambda_{ij}(x_j-\mu_j) \tag{94}$$

where $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$ is the precision matrix. When we substitute $\mathbf{\Lambda} = \mathbf{\Lambda}^{\mathrm{A}} + \mathbf{\Lambda}^{\mathrm{S}}$ into (94) we see that the term involving $\mathbf{\Lambda}^{\mathrm{A}}$ vanishes since for every positive term there is an equal and opposite negative term. Thus we can always take $\mathbf{\Lambda}$ to be symmetric.

2.18 We start by pre-multiplying both sides of (2.45) by $\mathbf{u}_i^\dagger$, the conjugate transpose of $\mathbf{u}_i$. This gives us

$$\mathbf{u}_i^\dagger\mathbf{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i^\dagger\mathbf{u}_i. \tag{95}$$

Next consider the conjugate transpose of (2.45) and post-multiply it by $\mathbf{u}_i$, which gives us

$$\mathbf{u}_i^\dagger\mathbf{\Sigma}^\dagger\mathbf{u}_i = \lambda_i^*\mathbf{u}_i^\dagger\mathbf{u}_i. \tag{96}$$

where λ_i^* is the complex conjugate of λ_i. We now subtract (95) from (96) and use the fact the $\mathbf{\Sigma}$ is real and symmetric and hence $\mathbf{\Sigma} = \mathbf{\Sigma}^\dagger$, to get

$$0 = (\lambda_i^* - \lambda_i)\mathbf{u}_i^\dagger\mathbf{u}_i.$$

Hence $\lambda_i^* = \lambda_i$ and so λ_i must be real.

Now consider

$$\begin{aligned}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j\lambda_j &= \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_j \\ &= \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{u}_j \\ &= (\boldsymbol{\Sigma}\mathbf{u}_i)^{\mathrm{T}}\,\mathbf{u}_j \\ &= \lambda_i\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j,\end{aligned}$$

where we have used (2.45) and the fact that $\boldsymbol{\Sigma}$ is symmetric. If we assume that $0 \neq \lambda_i \neq \lambda_j \neq 0$, the only solution to this equation is that $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = 0$, i.e., that $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthogonal.

If $0 \neq \lambda_i = \lambda_j \neq 0$, any linear combination of $\mathbf{u}_i$ and $\mathbf{u}_j$ will be an eigenvector with eigenvalue $\lambda = \lambda_i = \lambda_j$, since, from (2.45),

$$\begin{aligned}\boldsymbol{\Sigma}(a\mathbf{u}_i + b\mathbf{u}_j) &= a\lambda_i\mathbf{u}_i + b\lambda_j\mathbf{u}_j \\ &= \lambda(a\mathbf{u}_i + b\mathbf{u}_j).\end{aligned}$$

Assuming that $\mathbf{u}_i \neq \mathbf{u}_j$, we can construct

$$\begin{aligned}\mathbf{u}_\alpha &= a\mathbf{u}_i + b\mathbf{u}_j \\ \mathbf{u}_\beta &= c\mathbf{u}_i + d\mathbf{u}_j\end{aligned}$$

such that $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$ are mutually orthogonal and of unit length. Since $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthogonal to $\mathbf{u}_k$ ($k \neq i$, $k \neq j$), so are $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$. Thus, $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$ satisfy (2.46).

Finally, if $\lambda_i = 0$, $\boldsymbol{\Sigma}$ must be singular, with $\mathbf{u}_i$ lying in the nullspace of $\boldsymbol{\Sigma}$. In this case, $\mathbf{u}_i$ will be orthogonal to the eigenvectors projecting onto the rowspace of $\boldsymbol{\Sigma}$ and we can chose $\|\mathbf{u}_i\| = 1$, so that (2.46) is satisfied. If more than one eigenvalue equals zero, we can chose the corresponding eigenvectors arbitrily, as long as they remain in the nullspace of $\boldsymbol{\Sigma}$, and so we can chose them to satisfy (2.46).

2.19 We can write the r.h.s. of (2.48) in matrix form as

$$\sum_{i=1}^{D}\lambda_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} = \mathbf{M},$$

where $\mathbf{U}$ is a $D \times D$ matrix with the eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_D$ as its columns and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_D$ along its diagonal.

Thus we have

$$\mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\mathbf{U} = \boldsymbol{\Lambda}.$$

However, from (2.45)–(2.47), we also have that

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

and so $\mathbf{M} = \boldsymbol{\Sigma}$ and (2.48) holds.

Moreover, since $\mathbf{U}$ is orthonormal, $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$ and so

$$\boldsymbol{\Sigma}^{-1} = \left(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\right)^{-1} = \left(\mathbf{U}^{\mathrm{T}}\right)^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}} = \sum_{i=1}^{D}\lambda_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}.$$

2.20 Since $\mathbf{u}_1, \ldots, \mathbf{u}_D$ constitute a basis for $\mathbb{R}^D$, we can write

$$\mathbf{a} = \hat{a}_1\mathbf{u}_1 + \hat{a}_2\mathbf{u}_2 + \ldots + \hat{a}_D\mathbf{u}_D,$$

where $\hat{a}_1, \ldots, \hat{a}_D$ are coefficients obtained by projecting $\mathbf{a}$ on $\mathbf{u}_1, \ldots, \mathbf{u}_D$. Note that they typically do *not* equal the elements of $\mathbf{a}$.

Using this we can write

$$\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a} = \left(\hat{a}_1\mathbf{u}_1^{\mathrm{T}} + \ldots + \hat{a}_D\mathbf{u}_D^{\mathrm{T}}\right)\boldsymbol{\Sigma}\left(\hat{a}_1\mathbf{u}_1 + \ldots + \hat{a}_D\mathbf{u}_D\right)$$

and combining this result with (2.45) we get

$$\left(\hat{a}_1\mathbf{u}_1^{\mathrm{T}} + \ldots + \hat{a}_D\mathbf{u}_D^{\mathrm{T}}\right)\left(\hat{a}_1\lambda_1\mathbf{u}_1 + \ldots + \hat{a}_D\lambda_D\mathbf{u}_D\right).$$

Now, since $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = 1$ only if $i = j$, and 0 otherwise, this becomes

$$\hat{a}_1^2\lambda_1 + \ldots + \hat{a}_D^2\lambda_D$$

and since $\mathbf{a}$ is real, we see that this expression will be strictly positive for any non-zero $\mathbf{a}$, if all eigenvalues are strictly positive. It is also clear that if an eigenvalue, λ_i, is zero or negative, there exist a vector $\mathbf{a}$ (e.g. $\mathbf{a} = \mathbf{u}_i$), for which this expression will be less than or equal to zero. Thus, that a matrix has eigenvectors which are all strictly positive is a sufficient and necessary condition for the matrix to be positive definite.

2.21 A $D \times D$ matrix has D^2 elements. If it is symmetric then the elements not on the leading diagonal form pairs of equal value. There are D elements on the diagonal so the number of elements not on the diagonal is $D^2 - D$ and only half of these are independent giving

$$\frac{D^2 - D}{2}.$$

If we now add back the D elements on the diagonal we get

$$\frac{D^2 - D}{2} + D = \frac{D(D+1)}{2}.$$

2.22 Consider a matrix $\mathbf{M}$ which is symmetric, so that $\mathbf{M}^{\mathrm{T}} = \mathbf{M}$. The inverse matrix $\mathbf{M}^{-1}$ satisfies

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}.$$

Taking the transpose of both sides of this equation, and using the relation (C.1), we obtain

$$\left(\mathbf{M}^{-1}\right)^{\mathrm{T}}\mathbf{M}^{\mathrm{T}} = \mathbf{I}^{\mathrm{T}} = \mathbf{I}$$

since the identity matrix is symmetric. Making use of the symmetry condition for $\mathbf{M}$ we then have

$$\left(\mathbf{M}^{-1}\right)^{\mathrm{T}}\mathbf{M} = \mathbf{I}$$

and hence, from the definition of the matrix inverse,

$$\left(\mathbf{M}^{-1}\right)^{\mathrm{T}} = \mathbf{M}^{-1}$$

and so $\mathbf{M}^{-1}$ is also a symmetric matrix.

2.23 Recall that the transformation (2.51) diagonalizes the coordinate system and that the quadratic form (2.44), corresponding to the square of the Mahalanobis distance, is then given by (2.50). This corresponds to a shift in the origin of the coordinate system and a rotation so that the hyper-ellipsoidal contours along which the Mahalanobis distance is constant become axis aligned. The volume contained within any one such contour is unchanged by shifts and rotations. We now make the further transformation $z_i = \lambda_i^{1/2} y_i$ for $i = 1, \ldots, D$. The volume within the hyper-ellipsoid then becomes

$$\int \prod_{i=1}^{D} \mathrm{d}y_i = \prod_{i=1}^{D} \lambda_i^{1/2} \int \prod_{i=1}^{D} \mathrm{d}z_i = |\mathbf{\Sigma}|^{1/2} V_D \Delta^D$$

where we have used the property that the determinant of $\mathbf{\Sigma}$ is given by the product of its eigenvalues, together with the fact that in the z coordinates the volume has become a sphere of radius Δ whose volume is $V_D \Delta^D$.

2.24 Multiplying the left hand side of (2.76) by the matrix (2.287) trivially gives the identity matrix. On the right hand side consider the four blocks of the resulting partitioned matrix:

upper left

$$\mathbf{AM} - \mathbf{BD}^{-1}\mathbf{CM} = (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1} = \mathbf{I}$$

upper right

$$\begin{aligned} -\mathbf{AMBD}^{-1} + \mathbf{BD}^{-1} + \mathbf{BD}^{-1}\mathbf{CMBD}^{-1} \\ = \quad -(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1} + \mathbf{BD}^{-1} \\ = \quad -\mathbf{BD}^{-1} + \mathbf{BD}^{-1} = \mathbf{0} \end{aligned}$$

lower left

$$\mathbf{CM} - \mathbf{DD}^{-1}\mathbf{CM} = \mathbf{CM} - \mathbf{CM} = \mathbf{0}$$

lower right

$$-\mathbf{CMBD}^{-1} + \mathbf{DD}^{-1} + \mathbf{DD}^{-1}\mathbf{CMBD}^{-1} = \mathbf{DD}^{-1} = \mathbf{I}.$$

Thus the right hand side also equals the identity matrix.

2.25 We first of all take the joint distribution $p(\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c)$ and marginalize to obtain the distribution $p(\mathbf{x}_a, \mathbf{x}_b)$. Using the results of Section 2.3.2 this is again a Gaussian distribution with mean and covariance given by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}.$$

From Section 2.3.1 the distribution $p(\mathbf{x}_a, \mathbf{x}_b)$ is then Gaussian with mean and covariance given by (2.81) and (2.82) respectively.

2.26 Multiplying the left hand side of (2.289) by $(\mathbf{A} + \mathbf{BCD})$ trivially gives the identity matrix $\mathbf{I}$. On the right hand side we obtain

$$\begin{aligned}
&(\mathbf{A} + \mathbf{BCD})(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}) \\
&\quad = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&\qquad -\mathbf{BCDA}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&\quad = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{BC}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&\quad = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{BCDA}^{-1} = \mathbf{I}
\end{aligned}$$

2.27 From $\mathbf{y} = \mathbf{x} + \mathbf{z}$ we have trivially that $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{z}]$. For the covariance we have

$$\begin{aligned}
\mathrm{cov}[\mathbf{y}] &= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}] + \mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{x} - \mathbb{E}[\mathbf{x}] + \mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] + \mathbb{E}\left[(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}\right] \\
&\quad + \underbrace{\mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}\right]}_{=0} + \underbrace{\mathbb{E}\left[(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right]}_{=0} \\
&= \mathrm{cov}[\mathbf{x}] + \mathrm{cov}[\mathbf{z}]
\end{aligned}$$

where we have used the independence of $\mathbf{x}$ and $\mathbf{z}$, together with $\mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])] = \mathbb{E}[(\mathbf{z} - \mathbb{E}[\mathbf{z}])] = 0$, to set the third and fourth terms in the expansion to zero. For 1-dimensional variables the covariances become variances and we obtain the result of Exercise 1.10 as a special case.

2.28 For the marginal distribution $p(\mathbf{x})$ we see from (2.92) that the mean is given by the upper partition of (2.108) which is simply $\boldsymbol{\mu}$. Similarly from (2.93) we see that the covariance is given by the top left partition of (2.105) and is therefore given by $\boldsymbol{\Lambda}^{-1}$.

Now consider the conditional distribution $p(\mathbf{y}|\mathbf{x})$. Applying the result (2.81) for the conditional mean we obtain

$$\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{Ax} + \mathbf{b}.$$

Similarly applying the result (2.82) for the covariance of the conditional distribution we have

$$\mathrm{cov}[\mathbf{y}|\mathbf{x}] = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} = \mathbf{L}^{-1}$$

as required.

2.29 We first define

$$\mathbf{X} = \boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} \tag{97}$$

and

$$\mathbf{W} = -\mathbf{L}\mathbf{A}, \quad \text{and thus} \quad \mathbf{W}^{\mathrm{T}} = -\mathbf{A}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}} = -\mathbf{A}^{\mathrm{T}}\mathbf{L}, \tag{98}$$

since $\mathbf{L}$ is symmetric. We can use (97) and (98) to re-write (2.104) as

$$\mathbf{R} = \begin{pmatrix} \mathbf{X} & \mathbf{W}^{\mathrm{T}} \\ \mathbf{W} & \mathbf{L} \end{pmatrix}$$

and using (2.76) we get

$$\begin{pmatrix} \mathbf{X} & \mathbf{W}^{\mathrm{T}} \\ \mathbf{W} & \mathbf{L} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} \\ -\mathbf{L}^{-1}\mathbf{W}\mathbf{M} & \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{W}\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} \end{pmatrix}$$

where now

$$\mathbf{M} = \left(\mathbf{X} - \mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{W}\right)^{-1}.$$

Substituting $\mathbf{X}$ and $\mathbf{W}$ using (97) and (98), respectively, we get

$$\mathbf{M} = \left(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1}\mathbf{L}\mathbf{A}\right)^{-1} = \boldsymbol{\Lambda}^{-1},$$

$$-\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} = \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1} = \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}$$

and

$$\begin{aligned} \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{W}\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} &= \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{L}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1} \\ &= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}, \end{aligned}$$

as required.

2.30 Substituting the leftmost expression of (2.105) for $\mathbf{R}^{-1}$ in (2.107), we get

$$\begin{aligned} &\begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{S}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \\ \mathbf{S}\mathbf{b} \end{pmatrix} \\ &\quad = \begin{pmatrix} \boldsymbol{\Lambda}^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b}\right) + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b}\right) + \left(\mathbf{S}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right)\mathbf{S}\mathbf{b} \end{pmatrix} \\ &\quad = \begin{pmatrix} \boldsymbol{\mu} - \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \\ \mathbf{A}\boldsymbol{\mu} - \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} + \mathbf{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \end{pmatrix} \\ &\quad = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} - \mathbf{b} \end{pmatrix} \end{aligned}$$

2.31 Since $\mathbf{y} = \mathbf{x} + \mathbf{z}$ we can write the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ in the form $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_\mathbf{z} + \mathbf{x}, \boldsymbol{\Sigma}_\mathbf{z})$. This gives a decomposition of the joint distribution of $\mathbf{x}$ and $\mathbf{y}$ in the form $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ where $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x})$. This therefore takes the form of (2.99) and (2.100) in which we can identify $\boldsymbol{\mu} \to \boldsymbol{\mu}_\mathbf{x}$, $\boldsymbol{\Lambda}^{-1} \to \boldsymbol{\Sigma}_\mathbf{x}$, $\mathbf{A} \to \mathbf{I}$, $\mathbf{b} \to \boldsymbol{\mu}_\mathbf{z}$ and $\mathbf{L}^{-1} \to \boldsymbol{\Sigma}_\mathbf{z}$. We can now obtain the marginal distribution $p(\mathbf{y})$ by making use of the result (2.115) from which we obtain $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_\mathbf{x} + \boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z} + \boldsymbol{\Sigma}_\mathbf{x})$. Thus both the means and the covariances are additive, in agreement with the results of Exercise 2.27.

2.32 The quadratic form in the exponential of the joint distribution is given by

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathrm{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^\mathrm{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}). \tag{99}$$

We now extract all of those terms involving $\mathbf{x}$ and assemble them into a standard Gaussian quadratic form by completing the square

$$\begin{aligned} &= -\frac{1}{2}\mathbf{x}^\mathrm{T}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})\mathbf{x} + \mathbf{x}^\mathrm{T}\left[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\mathrm{T}\mathbf{L}(\mathbf{y} - \mathbf{b})\right] + \text{const} \\ &= -\frac{1}{2}(\mathbf{x} - \mathbf{m})^\mathrm{T}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})(\mathbf{x} - \mathbf{m}) \\ &\quad + \frac{1}{2}\mathbf{m}^\mathrm{T}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})\mathbf{m} + \text{const} \end{aligned} \tag{100}$$

where

$$\mathbf{m} = (\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\left[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\mathrm{T}\mathbf{L}(\mathbf{y} - \mathbf{b})\right].$$

We can now perform the integration over $\mathbf{x}$ which eliminates the first term in (100). Then we extract the terms in $\mathbf{y}$ from the final term in (100) and combine these with the remaining terms from the quadratic form (99) which depend on $\mathbf{y}$ to give

$$\begin{aligned} &= -\frac{1}{2}\mathbf{y}^\mathrm{T}\left\{\mathbf{L} - \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{L}\right\}\mathbf{y} \\ &\quad + \mathbf{y}^\mathrm{T}\left[\left\{\mathbf{L} - \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{L}\right\}\mathbf{b} \right. \\ &\quad \left. + \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\right]. \end{aligned} \tag{101}$$

We can identify the precision of the marginal distribution $p(\mathbf{y})$ from the second order term in $\mathbf{y}$. To find the corresponding covariance, we take the inverse of the precision and apply the Woodbury inversion formula (2.289) to give

$$\left\{\mathbf{L} - \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{L}\right\}^{-1} = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathrm{T} \tag{102}$$

which corresponds to (2.110).

Next we identify the mean $\boldsymbol{\nu}$ of the marginal distribution. To do this we make use of (102) in (101) and then complete the square to give

$$-\frac{1}{2}(\mathbf{y} - \boldsymbol{\nu})^\mathrm{T}\left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathrm{T}\right)^{-1}(\mathbf{y} - \boldsymbol{\nu}) + \text{const}$$

where

$$\boldsymbol{\nu} = \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right)\left[(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{b} + \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\right].$$

Now consider the two terms in the square brackets, the first one involving $\mathbf{b}$ and the second involving $\boldsymbol{\mu}$. The first of these contribution simply gives $\mathbf{b}$, while the term in $\boldsymbol{\mu}$ can be written

$$\begin{aligned} &= \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right)\mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} \\ &= \mathbf{A}(\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})(\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\mu} \end{aligned}$$

where we have used the general result $(\mathbf{B}\mathbf{C})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}$. Hence we obtain (2.109).

2.33 To find the conditional distribution $p(\mathbf{x}|\mathbf{y})$ we start from the quadratic form (99) corresponding to the joint distribution $p(\mathbf{x}, \mathbf{y})$. Now, however, we treat $\mathbf{y}$ as a constant and simply complete the square over $\mathbf{x}$ to give

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) \\ = & -\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x} + \mathbf{x}^{\mathrm{T}}\left\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}\mathbf{L}(\mathbf{y} - \mathbf{b})\right\} + \mathrm{const} \\ = & -\frac{1}{2}(\mathbf{x} - \mathbf{m})^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})(\mathbf{x} - \mathbf{m}) \end{aligned}$$

where, as in the solution to Exercise 2.32, we have defined

$$\mathbf{m} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\left\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b})\right\}$$

from which we obtain directly the mean and covariance of the conditional distribution in the form (2.111) and (2.112).

2.34 Differentiating (2.118) with respect to $\boldsymbol{\Sigma}$ we obtain two terms:

$$-\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}).$$

For the first term, we can apply (C.28) directly to get

$$-\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\ln|\boldsymbol{\Sigma}| = -\frac{N}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathrm{T}} = -\frac{N}{2}\boldsymbol{\Sigma}^{-1}.$$

For the second term, we first re-write the sum

$$\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = N\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right],$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}.$$

Using this together with (C.21), in which $x = \Sigma_{ij}$ (element (i, j) in $\boldsymbol{\Sigma}$), and properties of the trace we get

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_{ij}}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) &= N\frac{\partial}{\partial \Sigma_{ij}}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \\
&= N\mathrm{Tr}\left[\frac{\partial}{\partial \Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \\
&= -N\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \\
&= -N\mathrm{Tr}\left[\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\right] \\
&= -N\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\right)_{ij}
\end{aligned}$$

where we have used (C.26). Note that in the last step we have ignored the fact that $\Sigma_{ij} = \Sigma_{ji}$, so that $\partial\boldsymbol{\Sigma}/\partial\Sigma_{ij}$ has a 1 in position (i, j) only and 0 everywhere else. Treating this result as valid nevertheless, we get

$$-\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = \frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}.$$

Combining the derivatives of the two terms and setting the result to zero, we obtain

$$\frac{N}{2}\boldsymbol{\Sigma}^{-1} = \frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}.$$

Re-arrangement then yields

$$\boldsymbol{\Sigma} = \mathbf{S}$$

as required.

2.35 **NOTE**: In PRML, this exercise contains a typographical error; $\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_m\right]$ should be $\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_m^{\mathrm{T}}\right]$ on the l.h.s. of (2.291).

The derivation of (2.62) is detailed in the text between (2.59) (page 82) and (2.62) (page 83).

If $m = n$ then, using (2.62) we have $\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$, whereas if $n \neq m$ then the two data points $\mathbf{x}_n$ and $\mathbf{x}_m$ are independent and hence $\mathbb{E}[\mathbf{x}_n\mathbf{x}_m] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$ where we have used (2.59). Combining these results we obtain (2.291). From (2.59) and

(2.62) we then have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\Sigma}_{\mathrm{ML}}\right] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(\mathbf{x}_n - \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m\right)\left(\mathbf{x}_n^{\mathrm{T}} - \frac{1}{N}\sum_{l=1}^{N}\mathbf{x}_l^{\mathrm{T}}\right)\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \frac{2}{N}\mathbf{x}_n\sum_{m=1}^{N}\mathbf{x}_m^{\mathrm{T}} + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}\mathbf{x}_m\mathbf{x}_l^{\mathrm{T}}\right] \\
&= \left\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} - 2\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right\} \\
&= \left(\frac{N-1}{N}\right)\boldsymbol{\Sigma} \qquad (103)
\end{aligned}$$

as required.

2.36 **NOTE**: In the 1st printing of PRML, there are mistakes that affect this solution. The sign in (2.129) is incorrect, and this equation should read

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1}z(\theta^{(N-1)}).$$

Then, in order to be consistent with the assumption that $f(\theta) > 0$ for $\theta > \theta^\star$ and $f(\theta) < 0$ for $\theta < \theta^\star$ in Figure 2.10, we should find the root of the expected *negative* log likelihood. This lead to sign changes in (2.133) and (2.134), but in (2.135), these are cancelled against the change of sign in (2.129), so in effect, (2.135) remains unchanged. Also, $\mathbf{x}_n$ should be x_n on the l.h.s. of (2.133). Finally, the labels μ and μ_{ML} in Figure 2.11 should be interchanged and there are corresponding changes to the caption (see errata on the PRML web site for details).

Consider the expression for $\sigma^2_{(N)}$ and separate out the contribution from observation x_N to give

$$\begin{aligned}
\sigma^2_{(N)} &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N-1}(x_n - \mu)^2 + \frac{(x_N - \mu)^2}{N} \\
&= \frac{N-1}{N}\sigma^2_{(N-1)} + \frac{(x_N - \mu)^2}{N} \\
&= \sigma^2_{(N-1)} - \frac{1}{N}\sigma^2_{(N-1)} + \frac{(x_N - \mu)^2}{N} \\
&= \sigma^2_{(N-1)} + \frac{1}{N}\left\{(x_N - \mu)^2 - \sigma^2_{(N-1)}\right\}. \qquad (104)
\end{aligned}$$

If we substitute the expression for a Gaussian distribution into the result (2.135) for

the Robbins-Monro procedure applied to maximizing likelihood, we obtain

$$\begin{aligned}
\sigma^2_{(N)} &= \sigma^2_{(N-1)} + a_{N-1}\frac{\partial}{\partial\sigma^2_{(N-1)}}\left\{-\frac{1}{2}\ln\sigma^2_{(N-1)} - \frac{(x_N-\mu)^2}{2\sigma^2_{(N-1)}}\right\} \\
&= \sigma^2_{(N-1)} + a_{N-1}\left\{-\frac{1}{2\sigma^2_{(N-1)}} + \frac{(x_N-\mu)^2}{2\sigma^4_{(N-1)}}\right\} \\
&= \sigma^2_{(N-1)} + \frac{a_{N-1}}{2\sigma^4_{(N-1)}}\left\{(x_N-\mu)^2 - \sigma^2_{(N-1)}\right\}. \qquad (105)
\end{aligned}$$

Comparison of (105) with (104) allows us to identify

$$a_{N-1} = \frac{2\sigma^4_{(N-1)}}{N}.$$

2.37 **NOTE**: In PRML, this exercise requires the additional assumption that we can use the known true mean, $\boldsymbol{\mu}$, in (2.122). Furthermore, for the derivation of the Robbins-Monro sequential estimation formula, we assume that the covariance matrix is restricted to be diagonal. Starting from (2.122), we have

$$\begin{aligned}
\boldsymbol{\Sigma}^{(N)}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} \\
&= \frac{1}{N}\sum_{n=1}^{N-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} \\
&\quad + \frac{1}{N}(\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} \\
&= \frac{N-1}{N}\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}} + \frac{1}{N}(\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} \\
&= \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}} + \frac{1}{N}\left((\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} - \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right). \qquad (106)
\end{aligned}$$

From Solution 2.34, we know that

$$\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}}\ln p(\mathbf{x}_N|\boldsymbol{\mu},\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}) \\
&= \frac{1}{2}\left(\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)^{-1}\left((\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} - \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)\left(\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)^{-1} \\
&= \frac{1}{2}\left(\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)^{-2}\left((\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} - \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)
\end{aligned}$$

where we have used the assumption that $\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}$, and hence $\left(\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)^{-1}$, is diag-

onal. If we substitute this into the multivariate form of (2.135), we get

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} + \mathbf{A}_{N-1}\frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right)^{-2}\left(\left(\mathbf{x}_N - \boldsymbol{\mu}\right)\left(\mathbf{x}_N - \boldsymbol{\mu}\right)^{\mathrm{T}} - \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right) \tag{107}$$

where $\mathbf{A}_{N-1}$ is a matrix of coefficients corresponding to a_{N-1} in (2.135). By comparing (106) with (107), we see that if we choose

$$\mathbf{A}_{N-1} = \frac{2}{N}\left(\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right)^{2}.$$

we recover (106). Note that if the covariance matrix was restricted further, to the form $\sigma^2\mathbf{I}$, i.e. a spherical Gaussian, the coefficient in (107) would again become a scalar.

2.38 The exponent in the posterior distribution of (2.140) takes the form

$$-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 = -\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right) + \mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_{n=1}^{N}x_n\right) + \text{const.}$$

where 'const.' denotes terms independent of μ. Following the discussion of (2.71) we see that the variance of the posterior distribution is given by

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}.$$

Similarly the mean is given by

$$\begin{aligned}\mu_N &= \left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_{n=1}^{N}x_n\right) \\ &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}. \qquad (108)\end{aligned}$$

(109)

2.39 From (2.142), we see directly that

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{N-1}{\sigma^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}. \tag{110}$$

We also note for later use, that

$$\frac{1}{\sigma_N^2} = \frac{\sigma^2+N\sigma_0^2}{\sigma_0^2\sigma^2} = \frac{\sigma^2+\sigma_{N-1}^2}{\sigma_{N-1}^2\sigma^2} \tag{111}$$

and similarly

$$\frac{1}{\sigma_{N-1}^2} = \frac{\sigma^2 + (N-1)\sigma_0^2}{\sigma_0^2\sigma^2}. \tag{112}$$

Using (2.143), we can rewrite (2.141) as

$$\begin{aligned}\mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{\sigma_0^2 \sum_{n=1}^{N} x_n}{N\sigma_0^2 + \sigma^2} \\ &= \frac{\sigma^2\mu_0 + \sigma_0^2 \sum_{n=1}^{N-1} x_n}{N\sigma_0^2 + \sigma^2} + \frac{\sigma_0^2 x_N}{N\sigma_0^2 + \sigma^2}.\end{aligned}$$

Using (2.141), (111) and (112), we can rewrite the first term of this expression as

$$\frac{\sigma_N^2}{\sigma_{N-1}^2}\frac{\sigma^2\mu_0 + \sigma_0^2 \sum_{n=1}^{N-1} x_n}{(N-1)\sigma_0^2 + \sigma^2} = \frac{\sigma_N^2}{\sigma_{N-1}^2}\mu_{N-1}.$$

Similarly, using (111), the second term can be rewritten as

$$\frac{\sigma_N^2}{\sigma^2}x_N$$

and so

$$\mu_N = \frac{\sigma_N^2}{\sigma_{N-1}^2}\mu_{N-1} + \frac{\sigma_N^2}{\sigma^2}x_N. \tag{113}$$

Now consider

$$\begin{aligned}p(\mu|\mu_N, \sigma_N^2) &= p(\mu|\mu_{N-1}, \sigma_{N-1}^2)p(x_N|\mu, \sigma^2) \\ &= \mathcal{N}(\mu|\mu_{N-1}, \sigma_{N-1}^2)\mathcal{N}(x_N|\mu, \sigma^2) \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{\mu_{N-1}^2 - 2\mu\mu_{N-1} + \mu^2}{\sigma_{N-1}^2} + \frac{x_N^2 - 2x_N\mu + \mu^2}{\sigma^2}\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{\sigma^2(\mu_{N-1}^2 - 2\mu\mu_{N-1} + \mu^2)}{\sigma_{N-1}^2\sigma^2}\right.\right. \\ &\qquad \left.\left. + \frac{\sigma_{N-1}^2(x_N^2 - 2x_N\mu + \mu^2)}{\sigma_{N-1}^2\sigma^2}\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\frac{(\sigma_{N-1}^2 + \sigma^2)\mu^2 - 2(\sigma^2\mu_{N-1} + \sigma_{N-1}^2 x_N)\mu}{\sigma_{N-1}^2\sigma^2}\right\} + C,\end{aligned}$$

where C accounts for all the remaining terms that are independent of μ. From this, we can directly read off

$$\frac{1}{\sigma_N^2} = \frac{\sigma^2 + \sigma_{N-1}^2}{\sigma_{N-1}^2\sigma^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}$$

and

$$\begin{aligned}\mu_N &= \frac{\sigma^2\mu_{N-1} + \sigma^2_{N-1}x_N}{\sigma^2_{N-1} + \sigma^2} \\ &= \frac{\sigma^2}{\sigma^2_{N-1} + \sigma^2}\mu_{N-1} + \frac{\sigma^2_{N-1}}{\sigma^2_{N-1} + \sigma^2}x_N \\ &= \frac{\sigma^2_N}{\sigma^2_{N-1}}\mu_{N-1} + \frac{\sigma^2_N}{\sigma^2}x_N\end{aligned}$$

and so we have recovered (110) and (113).

2.40 The posterior distribution is proportional to the product of the prior and the likelihood function

$$p(\boldsymbol{\mu}|\mathbf{X}) \propto p(\boldsymbol{\mu})\prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Thus the posterior is proportional to an exponential of a quadratic form in $\boldsymbol{\mu}$ given by

$$\begin{aligned}&-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \\ &= -\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\mu} + \boldsymbol{\mu}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\mathbf{x}_n\right) + \text{const}\end{aligned}$$

where 'const.' denotes terms independent of $\boldsymbol{\mu}$. Using the discussion following (2.71) we see that the mean and covariance of the posterior distribution are given by

$$\boldsymbol{\mu}_N = \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}N\boldsymbol{\mu}_{\mathrm{ML}}\right) \tag{114}$$

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1} \tag{115}$$

where $\boldsymbol{\mu}_{\mathrm{ML}}$ is the maximum likelihood solution for the mean given by

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n.$$

2.41 If we consider the integral of the Gamma distribution over τ and make the change of variable $b\tau = u$ we have

$$\begin{aligned}\int_0^\infty \mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau &= \frac{1}{\Gamma(a)}\int_0^\infty b^a\tau^{a-1}\exp(-b\tau)\,\mathrm{d}\tau \\ &= \frac{1}{\Gamma(a)}\int_0^\infty b^a u^{a-1}\exp(-u)b^{1-a}b^{-1}\,\mathrm{d}u \\ &= 1\end{aligned}$$

where we have used the definition (1.141) of the Gamma function.

2.42 We can use the same change of variable as in the previous exercise to evaluate the mean of the Gamma distribution

$$\begin{aligned}\mathbb{E}[\tau] &= \frac{1}{\Gamma(a)}\int_0^\infty b^a \tau^{a-1}\tau \exp(-b\tau)\,\mathrm{d}\tau \\ &= \frac{1}{\Gamma(a)}\int_0^\infty b^a u^a \exp(-u) b^{-a} b^{-1}\,\mathrm{d}u \\ &= \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b}\end{aligned}$$

where we have used the recurrence relation $\Gamma(a+1) = a\Gamma(a)$ for the Gamma function. Similarly we can find the variance by first evaluating

$$\begin{aligned}\mathbb{E}[\tau^2] &= \frac{1}{\Gamma(a)}\int_0^\infty b^a \tau^{a-1}\tau^2 \exp(-b\tau)\,\mathrm{d}\tau \\ &= \frac{1}{\Gamma(a)}\int_0^\infty b^a u^{a+1} \exp(-u) b^{-a-1} b^{-1}\,\mathrm{d}u \\ &= \frac{\Gamma(a+2)}{b^2\Gamma(a)} = \frac{(a+1)\Gamma(a+1)}{b^2\Gamma(a)} = \frac{a(a+1)}{b^2}\end{aligned}$$

and then using

$$\mathrm{var}[\tau] = \mathbb{E}[\tau^2] - \mathbb{E}[\tau]^2 = \frac{a(a+1)}{b^2} - \frac{a^2}{b^2} = \frac{a}{b^2}.$$

Finally, the mode of the Gamma distribution is obtained simply by differentiation

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left\{\tau^{a-1}\exp(-b\tau)\right\} = \left[\frac{a-1}{\tau} - b\right]\tau^{a-1}\exp(-b\tau) = 0$$

from which we obtain

$$\mathrm{mode}[\tau] = \frac{a-1}{b}.$$

Notice that the mode only exists if $a \geqslant 1$, since τ must be a non-negative quantity. This is also apparent in the plot of Figure 2.13.

2.43 To prove the normalization of the distribution (2.293) consider the integral

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{|x|^q}{2\sigma^2}\right)\mathrm{d}x = 2\int_0^\infty \exp\left(-\frac{x^q}{2\sigma^2}\right)\mathrm{d}x$$

and make the change of variable

$$u = \frac{x^q}{2\sigma^2}.$$

Using the definition (1.141) of the Gamma function, this gives

$$I = 2\int_0^\infty \frac{2\sigma^2}{q}(2\sigma^2 u)^{(1-q)/q}\exp(-u)\,\mathrm{d}u = \frac{2(2\sigma^2)^{1/q}\Gamma(1/q)}{q}$$

from which the normalization of (2.293) follows.

For the given noise distribution, the conditional distribution of the target variable given the input variable is

$$p(t|\mathbf{x},\mathbf{w},\sigma^2) = \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)}\exp\left(-\frac{|t-y(\mathbf{x},\mathbf{w})|^q}{2\sigma^2}\right).$$

The likelihood function is obtained by taking products of factors of this form, over all pairs $\{\mathbf{x}_n, t_n\}$. Taking the logarithm, and discarding additive constants, we obtain the desired result.

2.44 From Bayes' theorem we have

$$p(\mu,\lambda|\mathbf{X}) \propto p(\mathbf{X}|\mu,\lambda)p(\mu,\lambda),$$

where the factors on the r.h.s. are given by (2.152) and (2.154), respectively. Writing this out in full, we get

$$\begin{aligned} p(\mu,\lambda) \quad \propto \quad & \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^N x_n - \frac{\lambda}{2}\sum_{n=1}^N x_n^2\right\} \\ & (\beta\lambda)^{1/2}\exp\left[-\frac{\beta\lambda}{2}\left(\mu^2 - 2\mu\mu_0 + \mu_0^2\right)\right]\lambda^{a-1}\exp\left(-b\lambda\right), \end{aligned}$$

where we have used the defintions of the Gaussian and Gamma distributions and we have ommitted terms independent of μ and λ. We can rearrange this to obtain

$$\begin{aligned} & \lambda^{N/2}\lambda^{a-1}\exp\left\{-\left(b + \frac{1}{2}\sum_{n=1}^N x_n^2 + \frac{\beta}{2}\mu_0^2\right)\lambda\right\} \\ & \qquad (\lambda(N+\beta))^{1/2}\exp\left[-\frac{\lambda(N+\beta)}{2}\left(\mu^2 - \frac{2}{N+\beta}\left\{\beta\mu_0 + \sum_{n=1}^N x_n\right\}\mu\right)\right] \end{aligned}$$

and by completing the square in the argument of the second exponential,

$$\begin{aligned} & \lambda^{N/2}\lambda^{a-1}\exp\left\{-\left(b + \frac{1}{2}\sum_{n=1}^N x_n^2 + \frac{\beta}{2}\mu_0^2 - \frac{\left(\beta\mu_0 + \sum_{n=1}^N x_n\right)^2}{2(N+\beta)}\right)\lambda\right\} \\ & \qquad (\lambda(N+\beta))^{1/2}\exp\left[-\frac{\lambda(N+\beta)}{2}\left(\mu - \frac{\beta\mu_0 + \sum_{n=1}^N x_n}{N+\beta}\right)\right] \end{aligned}$$

we arrive at an (unnormalised) Gaussian-Gamma distribution,

$$\mathcal{N}\left(\mu|\mu_N,((N+\beta)\lambda)^{-1}\right)\mathrm{Gam}\left(\lambda|a_N,b_N\right),$$

with parameters

$$\begin{aligned}\mu_N &= \frac{\beta\mu_0+\sum_{n=1}^N x_n}{N+\beta}\\ a_N &= a+\frac{N}{2}\\ b_N &= b+\frac{1}{2}\sum_{n=1}^N x_n^2+\frac{\beta}{2}\mu_0^2-\frac{N+\beta}{2}\mu_N^2.\end{aligned}$$

2.45 We do this, as in the univariate case, by considering the likelihood function of $\mathbf{\Lambda}$ for a given data set $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$:

$$\begin{aligned}\prod_{n=1}^N\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu},\mathbf{\Lambda}^{-1}) &\propto |\mathbf{\Lambda}|^{N/2}\exp\left(-\frac{1}{2}\sum_{n=1}^N(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Lambda}(\mathbf{x}_n-\boldsymbol{\mu})\right)\\ &= |\mathbf{\Lambda}|^{N/2}\exp\left(-\frac{1}{2}\mathrm{Tr}\left[\mathbf{\Lambda}\mathbf{S}\right]\right),\end{aligned}$$

where $\mathbf{S}=\sum_n(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}$. By simply comparing with (2.155), we see that the functional dependence on $\mathbf{\Lambda}$ is indeed the same and thus a product of this likelihood and a Wishart prior will result in a Wishart posterior.

2.46 From (2.158), we have

$$\begin{aligned}&\int_0^\infty \frac{b^a e^{(-b\tau)}\tau^{a-1}}{\Gamma(a)}\left(\frac{\tau}{2\pi}\right)^{1/2}\exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\}\mathrm{d}\tau\\ &\qquad=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\int_0^\infty \tau^{a-1/2}\exp\left\{-\tau\left(b+\frac{(x-\mu)^2}{2}\right)\right\}\mathrm{d}\tau.\end{aligned}$$

We now make the proposed change of variable $z=\tau\Delta$, where $\Delta=b+(x-\mu)^2/2$, yielding

$$\begin{aligned}&\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\Delta^{-a-1/2}\int_0^\infty z^{a-1/2}\exp(-z)\,\mathrm{d}z\\ &\qquad=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\Delta^{-a-1/2}\Gamma(a+1/2)\end{aligned}$$

where we have used the definition of the Gamma function (1.141). Finally, we substitute $b + (x-\mu)^2/2$ for Δ, $\nu/2$ for a and $\nu/2\lambda$ for b:

$$\begin{aligned}
&\frac{\Gamma(-a+1/2)}{\Gamma(a)} b^a \left(\frac{1}{2\pi}\right)^{1/2} \Delta^{a-1/2} \\
&= \frac{\Gamma\left((\nu+1)/2\right)}{\Gamma(\nu/2)} \left(\frac{\nu}{2\lambda}\right)^{\nu/2} \left(\frac{1}{2\pi}\right)^{1/2} \left(\frac{\nu}{2\lambda} + \frac{(x-\mu)^2}{2}\right)^{-(\nu+1)/2} \\
&= \frac{\Gamma\left((\nu+1)/2\right)}{\Gamma(\nu/2)} \left(\frac{\nu}{2\lambda}\right)^{\nu/2} \left(\frac{1}{2\pi}\right)^{1/2} \left(\frac{\nu}{2\lambda}\right)^{-(\nu+1)/2} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-(\nu+1)/2} \\
&= \frac{\Gamma\left((\nu+1)/2\right)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\nu\pi}\right)^{1/2} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-(\nu+1)/2}
\end{aligned}$$

2.47 Ignoring the normalization constant, we write (2.159) as

$$\begin{aligned}
\mathrm{St}(x|\mu,\lambda,\nu) &\propto \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-(\nu-1)/2} \\
&= \exp\left(-\frac{\nu-1}{2} \ln\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]\right). \qquad (116)
\end{aligned}$$

For large ν, we make use of the Taylor expansion for the logarithm in the form

$$\ln(1+\epsilon) = \epsilon + O(\epsilon^2) \qquad (117)$$

to re-write (116) as

$$\begin{aligned}
&\exp\left(-\frac{\nu-1}{2} \ln\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]\right) \\
&= \exp\left(-\frac{\nu-1}{2}\left[\frac{\lambda(x-\mu)^2}{\nu} + O(\nu^{-2})\right]\right) \\
&= \exp\left(-\frac{\lambda(x-\mu)^2}{2} + O(\nu^{-1})\right).
\end{aligned}$$

We see that in the limit $\nu \to \infty$ this becomes, up to an overall constant, the same as a Gaussian distribution with mean μ and precision λ. Since the Student distribution is normalized to unity for all values of ν it follows that it must remain normalized in this limit. The normalization coefficient is given by the standard expression (2.42) for a univariate Gaussian.

2.48 Substituting expressions for the Gaussian and Gamma distributions into (2.161), we have

$$\begin{aligned}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \int_0^\infty \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\ &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\int_0^\infty \eta^{D/2}\eta^{\nu/2-1}e^{-\nu\eta/2}e^{-\eta\Delta^2/2}\,\mathrm{d}\eta.\end{aligned}$$

Now we make the change of variable

$$\tau = \eta\left[\frac{\nu}{2}+\frac{1}{2}\Delta^2\right]^{-1}$$

which gives

$$\begin{aligned}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\left[\frac{\nu}{2}+\frac{1}{2}\Delta^2\right]^{-D/2-\nu/2} \\ &\qquad \int_0^\infty \tau^{D/2+\nu/2-1}e^{-\tau}\,\mathrm{d}\tau \\ &= \frac{\Gamma(\nu/2+d/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}}\left[1+\frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2}\end{aligned}$$

as required.

The correct normalization of the multivariate Student's t-distribution follows directly from the fact that the Gaussian and Gamma distributions are normalized. From (2.161) we have

$$\begin{aligned}\int \mathrm{St}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu\right)\,\mathrm{d}\mathbf{x} &= \iint \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1}\right)\mathrm{Gam}\left(\eta|\nu/2,\nu/2\right)\,\mathrm{d}\eta\,\mathrm{d}\mathbf{x} \\ &= \iint \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1}\right)\,\mathrm{d}\mathbf{x}\,\mathrm{Gam}\left(\eta|\nu/2,\nu/2\right)\,\mathrm{d}\eta \\ &= \int \mathrm{Gam}\left(\eta|\nu/2,\nu/2\right)\,\mathrm{d}\eta = 1.\end{aligned}$$

2.49 If we make the change of variable $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$, we can write

$$\mathbb{E}[\mathbf{x}] = \int \mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)\mathbf{x}\,\mathrm{d}\mathbf{x} = \int \mathrm{St}(\mathbf{z}|\mathbf{0},\boldsymbol{\Lambda},\nu)(\mathbf{z}+\boldsymbol{\mu})\,\mathrm{d}\mathbf{z}.$$

In the factor $(\mathbf{z}+\boldsymbol{\mu})$ the first term vanishes as a consequence of the fact that the zero-mean Student distribution is an even function of $\mathbf{z}$ that is $\mathrm{St}(-\mathbf{z}|\mathbf{0},\boldsymbol{\Lambda},\nu) = \mathrm{St}(-\mathbf{z}|\mathbf{0},\boldsymbol{\Lambda},\nu)$. This leaves the second term, which equals $\boldsymbol{\mu}$ since the Student distribution is normalized.

The covariance of the multivariate Student can be re-expressed by using the expression for the multivariate Student distribution as a convolution of a Gaussian with a Gamma distribution given by (2.161) which gives

$$\begin{aligned}\mathrm{cov}[\mathbf{x}] &= \int \mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\mathbf{x} \\ &= \int_0^\infty \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\eta\boldsymbol{\Lambda})(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\mathbf{x}\,\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\ &= \int_0^\infty \eta^{-1}\boldsymbol{\Lambda}^{-1}\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta\end{aligned}$$

where we have used the standard result for the covariance of a multivariate Gaussian. We now substitute for the Gamma distribution using (2.146) to give

$$\begin{aligned}\mathrm{cov}[\mathbf{x}] &= \frac{1}{\Gamma(\nu/2)}\left(\frac{\nu}{2}\right)^{\nu/2}\int_0^\infty e^{-\nu\eta/2}\eta^{\nu/2-2}\,\mathrm{d}\eta\boldsymbol{\Lambda}^{-1} \\ &= \frac{\nu}{2}\frac{\Gamma(\nu/2-2)}{\Gamma(\nu/2)}\boldsymbol{\Lambda}^{-1} \\ &= \frac{\nu}{\nu-2}\boldsymbol{\Lambda}^{-1}\end{aligned}$$

where we have used the integral representation for the Gamma function, together with the standard result $\Gamma(1+x) = x\Gamma(x)$.

The mode of the Student distribution is obtained by differentiation

$$\nabla_{\mathbf{x}}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \frac{\Gamma(\nu/2+D/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}}\left[1+\frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2-1}\frac{1}{\nu}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}).$$

Provided $\boldsymbol{\Lambda}$ is non-singular we therefore obtain

$$\mathrm{mode}[\mathbf{x}] = \boldsymbol{\mu}.$$

2.50 Just like in univariate case (Exercise 2.47), we ignore the normalization coefficient, which leaves us with

$$\left[1+\frac{\Delta^2}{\nu}\right]^{-\nu/2-D/2} = \exp\left\{-\left(\frac{\nu}{2}+\frac{D}{2}\right)\ln\left[1+\frac{\Delta^2}{\nu}\right]\right\}$$

where Δ^2 is the squared Mahalanobis distance given by

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}).$$

Again we make use of (117) to give

$$\exp\left\{-\left(\frac{\nu}{2}+\frac{D}{2}\right)\ln\left[1+\frac{\Delta^2}{\nu}\right]\right\} = \exp\left\{-\frac{\Delta^2}{2}+O(1/\nu)\right\}.$$

As in the univariate case, in the limit $\nu\to\infty$ this becomes, up to an overall constant, the same as a Gaussian distribution, here with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$; the univariate normalization argument also applies in the multivariate case.

2.51 Using the relation (2.296) we have

$$1 = \exp(iA)\exp(-iA) = (\cos A + i\sin A)(\cos A - i\sin A) = \cos^2 A + \sin^2 A.$$

Similarly, we have

$$\begin{aligned}\cos(A-B) &= \Re\exp\{i(A-B)\} \\ &= \Re\exp(iA)\exp(-iB) \\ &= \Re(\cos A + i\sin A)(\cos B - i\sin B) \\ &= \cos A\cos B + \sin A\sin B.\end{aligned}$$

Finally

$$\begin{aligned}\sin(A-B) &= \Im\exp\{i(A-B)\} \\ &= \Im\exp(iA)\exp(-iB) \\ &= \Im(\cos A + i\sin A)(\cos B - i\sin B) \\ &= \sin A\cos B - \cos A\sin B.\end{aligned}$$

2.52 Expressed in terms of ξ the von Mises distribution becomes

$$p(\xi) \propto \exp\left\{m\cos(m^{-1/2}\xi)\right\}.$$

For large m we have $\cos(m^{-1/2}\xi) = 1 - m^{-1}\xi^2/2 + O(m^{-2})$ and so

$$p(\xi) \propto \exp\left\{-\xi^2/2\right\}$$

and hence $p(\theta) \propto \exp\{-m(\theta-\theta_0)^2/2\}$.

2.53 Using (2.183), we can write (2.182) as

$$\sum_{n=1}^{N}(\cos\theta_0\sin\theta_n - \cos\theta_n\sin\theta_0) = \cos\theta_0\sum_{n=1}^{N}\sin\theta_n - \sin\sum_{n=1}^{N}\cos\theta_n = 0.$$

Rearranging this, we get

$$\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n} = \frac{\sin\theta_0}{\cos\theta_0} = \tan\theta_0,$$

which we can solve w.r.t. θ_0 to obtain (2.184).

2.54 Differentiating the von Mises distribution (2.179) we have

$$p'(\theta) = -\frac{1}{2\pi I_0(m)}\exp\left\{m\cos(\theta-\theta_0)\right\}\sin(\theta-\theta_0)$$

which vanishes when $\theta = \theta_0$ or when $\theta = \theta_0 + \pi \pmod{2\pi}$. Differentiating again we have

$$p''(\theta) = -\frac{1}{2\pi I_0(m)}\exp\left\{m\cos(\theta-\theta_0)\right\}\left[\sin^2(\theta-\theta_0) + \cos(\theta-\theta_0)\right].$$

Since $I_0(m) > 0$ we see that $p''(\theta) < 0$ when $\theta = \theta_0$, which therefore represents a maximum of the density, while $p''(\theta) > 0$ when $\theta = \theta_0 + \pi \, (\mathrm{mod} 2\pi)$, which is therefore a minimum.

2.55 **NOTE**: In the 1st printing of PRML, equation (2.187), which will be the starting point for this solution, contains a typo. The "$-$" on the r.h.s. should be a "$+$", as is easily seen from (2.178) and (2.185).

From (2.169) and (2.184), we see that $\bar{\theta} = \theta_0^{\mathrm{ML}}$. Using this together with (2.168) and (2.177), we can rewrite (2.187) as follows:

$$\begin{aligned} A(m_{\mathrm{ML}}) &= \left(\frac{1}{N} \sum_{n=1}^{N} \cos\theta_n \right) \cos\theta_0^{\mathrm{ML}} + \left(\frac{1}{N} \sum_{n=1}^{N} \sin\theta_n \right) \sin\theta_0^{\mathrm{ML}} \\ &= \bar{r} \cos\bar{\theta} \cos\theta_0^{\mathrm{ML}} + \bar{r} \sin\bar{\theta} \sin\theta_0^{\mathrm{ML}} \\ &= \bar{r} \left(\cos^2\theta_0^{\mathrm{ML}} + \sin^2\theta_0^{\mathrm{ML}} \right) \\ &= \bar{r}. \end{aligned}$$

2.56 We can most conveniently cast distributions into standard exponential family form by taking the exponential of the logarithm of the distribution. For the Beta distribution (2.13) we have

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp\left\{ (a-1)\ln\mu + (b-1)\ln(1-\mu) \right\}$$

which we can identify as being in standard exponential form (2.194) with

$$h(\mu) = 1 \tag{118}$$

$$g(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{119}$$

$$\mathbf{u}(\mu) = \begin{pmatrix} \ln\mu \\ \ln(1-\mu) \end{pmatrix} \tag{120}$$

$$\boldsymbol{\eta}(a,b) = \begin{pmatrix} a-1 \\ b-1 \end{pmatrix}. \tag{121}$$

Applying the same approach to the gamma distribution (2.146) we obtain

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)} \exp\left\{ (a-1)\ln\lambda - b\lambda \right\}.$$

from which it follows that

$$h(\lambda) = 1 \tag{122}$$
$$g(a,b) = \frac{b^a}{\Gamma(a)} \tag{123}$$
$$\mathbf{u}(\lambda) = \begin{pmatrix} \lambda \\ \ln\lambda \end{pmatrix} \tag{124}$$
$$\boldsymbol{\eta}(a,b) = \begin{pmatrix} -b \\ a-1 \end{pmatrix}. \tag{125}$$

Finally, for the von Mises distribution (2.179) we make use of the identity (2.178) to give

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\{m\cos\theta\cos\theta_0 + m\sin\theta\sin\theta_0\}$$

from which we find

$$h(\theta) = 1 \tag{126}$$
$$g(\theta_0, m) = \frac{1}{2\pi I_0(m)} \tag{127}$$
$$\mathbf{u}(\theta) = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{128}$$
$$\boldsymbol{\eta}(\theta_0, m) = \begin{pmatrix} m\cos\theta_0 \\ m\sin\theta_0 \end{pmatrix}. \tag{129}$$

2.57 Starting from (2.43), we can rewrite the argument of the exponential as

$$-\frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

The last term is indepedent of $\mathbf{x}$ but depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and so should go into $g(\boldsymbol{\eta})$. The second term is already an inner product and can be kept as is. To deal with the first term, we define the D^2-dimensional vectors $\mathbf{z}$ and $\boldsymbol{\lambda}$, which consist of the columns of $\mathbf{x}\mathbf{x}^{\mathrm{T}}$ and $\boldsymbol{\Sigma}^{-1}$, respectively, stacked on top of each other. Now we can write the multivariate Gaussian distribution on the form (2.194), with

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\lambda} \end{bmatrix}$$
$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}$$
$$h(\mathbf{x}) = (2\pi)^{-D/2}$$
$$g(\boldsymbol{\eta}) = |\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right).$$

2.58 Taking the first derivative of (2.226) we obtain, as in the text,

$$-\nabla \ln g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

Taking the gradient again gives

$$\begin{aligned} -\nabla\nabla \ln g(\boldsymbol{\eta}) &= g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathrm{T}}\, \mathrm{d}\mathbf{x} \\ &\quad + \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &= \mathbb{E}[\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathrm{T}}] - \mathbb{E}[\mathbf{u}(\mathbf{x})]\mathbb{E}[\mathbf{u}(\mathbf{x})^{\mathrm{T}}] \\ &= \mathrm{cov}[\mathbf{u}(\mathbf{x})] \end{aligned}$$

where we have used the result (2.226).

2.59

$$\begin{aligned} \int \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \mathrm{d}x &= \frac{1}{\sigma} \int f(y) \frac{\mathrm{d}x}{\mathrm{d}y}\, \mathrm{d}y \\ &= \frac{1}{\sigma} \int f(y)\sigma\, \mathrm{d}y \\ &= \frac{\sigma}{\sigma} \int f(y)\, \mathrm{d}y = 1, \end{aligned}$$

since $f(x)$ integrates to 1.

2.60 The value of the density $p(\mathbf{x})$ at a point $\mathbf{x}_n$ is given by $h_{j(n)}$, where the notation $j(n)$ denotes that data point $\mathbf{x}_n$ falls within region j. Thus the log likelihood function takes the form

$$\sum_{n=1}^{N} \ln p(\mathbf{x}_n) = \sum_{n=1}^{N} \ln h_{j(n)}.$$

We now need to take account of the constraint that $p(\mathbf{x})$ must integrate to unity. Since $p(\mathbf{x})$ has the constant value h_i over region i, which has volume Δ_i, the normalization constraint becomes $\sum_i h_i \Delta_i = 1$. Introducing a Lagrange multiplier λ we then minimize the function

$$\sum_{n=1}^{N} \ln h_{j(n)} + \lambda \left(\sum_i h_i \Delta_i - 1 \right)$$

with respect to h_k to give

$$0 = \frac{n_k}{h_k} + \lambda \Delta_k$$

where n_k denotes the total number of data points falling within region k. Multiplying both sides by h_k, summing over k and making use of the normalization constraint,

we obtain $\lambda = -N$. Eliminating λ then gives our final result for the maximum likelihood solution for h_k in the form

$$h_k = \frac{n_k}{N}\frac{1}{\Delta_k}.$$

Note that, for equal sized bins $\Delta_k = \Delta$ we obtain a bin height h_k which is proportional to the fraction of points falling within that bin, as expected.

2.61 From (2.246) we have

$$p(\mathbf{x}) = \frac{K}{NV(\rho)}$$

where $V(\rho)$ is the volume of a D-dimensional hypersphere with radius ρ, where in turn ρ is the distance from $\mathbf{x}$ to its K^{th} nearest neighbour in the data set. Thus, in polar coordinates, if we consider sufficiently large values for the radial coordinate r, we have

$$p(\mathbf{x}) \propto r^{-D}.$$

If we consider the integral of $p(\mathbf{x})$ and note that the volume element $\mathrm{d}\mathbf{x}$ can be written as $r^{D-1}\,\mathrm{d}r$, we get

$$\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} \propto \int r^{-D} r^{D-1}\,\mathrm{d}r = \int r^{-1}\,\mathrm{d}r$$

which diverges logarithmically.

Chapter 3 Linear Models for Regression

3.1 **NOTE**: In the 1st printing of PRML, there is a 2 missing in the denominator of the argument to the 'tanh' function in equation (3.102).

Using (3.6), we have

$$\begin{aligned}
2\sigma(2a) - 1 &= \frac{2}{1+e^{-2a}} - 1 \\
&= \frac{2}{1+e^{-2a}} - \frac{1+e^{-2a}}{1+e^{-2a}} \\
&= \frac{1-e^{-2a}}{1+e^{-2a}} \\
&= \frac{e^{a}-e^{-a}}{e^{a}+e^{-a}} \\
&= \tanh(a)
\end{aligned}$$

If we now take $a_j = (x - \mu_j)/2s$, we can rewrite (3.101) as

$$\begin{aligned} y(\mathbf{x}, \mathbf{w}) &= w_0 + \sum_{j=1}^{M} w_j \sigma(2a_j) \\ &= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} \left(2\sigma(2a_j) - 1 + 1\right) \\ &= u_0 + \sum_{j=1}^{M} u_j \tanh(a_j), \end{aligned}$$

where $u_j = w_j/2$, for $j = 1, \ldots, M$, and $u_0 = w_0 + \sum_{j=1}^{M} w_j/2$.

3.2 We first write

$$\begin{aligned} \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{v} &= \boldsymbol{\Phi}\widetilde{\mathbf{v}} \\ &= \boldsymbol{\varphi}_1 \tilde{v}^{(1)} + \boldsymbol{\varphi}_2 \tilde{v}^{(2)} + \ldots + \boldsymbol{\varphi}_M \tilde{v}^{(M)} \end{aligned}$$

where $\boldsymbol{\varphi}_m$ is the m-th column of $\boldsymbol{\Phi}$ and $\widetilde{\mathbf{v}} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{v}$. By comparing this with the least squares solution in (3.15), we see that

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

corresponds to a projection of $\mathbf{t}$ onto the space spanned by the columns of $\boldsymbol{\Phi}$. To see that this is indeed an orthogonal projection, we first note that for any column of $\boldsymbol{\Phi}$, $\boldsymbol{\varphi}_j$,

$$\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\varphi}_j = \left[\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right]_j = \boldsymbol{\varphi}_j$$

and therefore

$$(\mathbf{y} - \mathbf{t})^{\mathrm{T}} \boldsymbol{\varphi}_j = (\boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} - \mathbf{t})^{\mathrm{T}} \boldsymbol{\varphi}_j = \mathbf{t}^{\mathrm{T}} \left(\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}} - \mathbf{I}\right)^{\mathrm{T}} \boldsymbol{\varphi}_j = 0$$

and thus $(\mathbf{y} - \mathbf{t})$ is ortogonal to every column of $\boldsymbol{\Phi}$ and hence is orthogonal to $\mathcal{S}$.

3.3 If we define $\mathbf{R} = \mathrm{diag}(r_1, \ldots, r_N)$ to be a diagonal matrix containing the weighting coefficients, then we can write the weighted sum-of-squares cost function in the form

$$E_D(\mathbf{w}) = \frac{1}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^{\mathrm{T}}\mathbf{R}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}).$$

Setting the derivative with respect to $\mathbf{w}$ to zero, and re-arranging, then gives

$$\mathbf{w}^{\star} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{t}$$

which reduces to the standard solution (3.15) for the case $\mathbf{R} = \mathbf{I}$.

If we compare (3.104) with (3.10)–(3.12), we see that r_n can be regarded as a precision (inverse variance) parameter, particular to the data point $(\mathbf{x}_n, t_n)$, that either replaces or scales β.

Alternatively, r_n can be regarded as an *effective* number of replicated observations of data point $(\mathbf{x}_n, t_n)$; this becomes particularly clear if we consider (3.104) with r_n taking positive integer values, although it is valid for any $r_n > 0$.

3.4 Let

$$\begin{aligned}\widetilde{y}_n &= w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) \\ &= y_n + \sum_{i=1}^{D} w_i \epsilon_{ni}\end{aligned}$$

where $y_n = y(x_n, \mathbf{w})$ and $\epsilon_{ni} \sim \mathcal{N}(0, \sigma^2)$ and we have used (3.105). From (3.106) we then define

$$\begin{aligned}\widetilde{E} &= \frac{1}{2}\sum_{n=1}^{N} \{\widetilde{y}_n - t_n\}^2 \\ &= \frac{1}{2}\sum_{n=1}^{N} \{\widetilde{y}_n^2 - 2\widetilde{y}_n t_n + t_n^2\} \\ &= \frac{1}{2}\sum_{n=1}^{N} \left\{ y_n^2 + 2y_n \sum_{i=1}^{D} w_i\epsilon_{ni} + \left(\sum_{i=1}^{D} w_i \epsilon_{ni}\right)^2 \right. \\ &\qquad \left. -2t_n y_n - 2t_n \sum_{i=1}^{D} w_i \epsilon_{ni} + t_n^2 \right\}.\end{aligned}$$

If we take the expectation of $\widetilde{E}$ under the distribution of ϵ_{ni}, we see that the second and fifth terms disappear, since $\mathbb{E}[\epsilon_{ni}] = 0$, while for the third term we get

$$\mathbb{E}\left[\left(\sum_{i=1}^{D} w_i \epsilon_{ni}\right)^2\right] = \sum_{i=1}^{D} w_i^2 \sigma^2$$

since the ϵ_{ni} are all independent with variance σ^2.

From this and (3.106) we see that

$$\mathbb{E}\left[\widetilde{E}\right] = E_D + \frac{1}{2}\sum_{i=1}^{D} w_i^2 \sigma^2,$$

as required.

3.5 We can rewrite (3.30) as

$$\frac{1}{2}\left(\sum_{j=1}^{M}|w_j|^q - \eta\right) \leqslant 0$$

where we have incorporated the $1/2$ scaling factor for convenience. Clearly this does not affect the constraint.

Employing the technique described in Appendix E, we can combine this with (3.12) to obtain the Lagrangian function

$$L(\mathbf{w},\lambda) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\left(\sum_{j=1}^{M}|w_j|^q - \eta\right)$$

and by comparing this with (3.29) we see immediately that they are identical in their dependence on $\mathbf{w}$.

Now suppose we choose a specific value of $\lambda > 0$ and minimize (3.29). Denoting the resulting value of $\mathbf{w}$ by $\mathbf{w}^\star(\lambda)$, and using the KKT condition (E.11), we see that the value of η is given by

$$\eta = \sum_{j=1}^{M}|w_j^\star(\lambda)|^q.$$

3.6 We first write down the log likelihood function which is given by

$$\ln L(\mathbf{W},\boldsymbol{\Sigma}) = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)).$$

First of all we set the derivative with respect to $\mathbf{W}$ equal to zero, giving

$$0 = -\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}.$$

Multiplying through by $\boldsymbol{\Sigma}$ and introducing the design matrix $\boldsymbol{\Phi}$ and the target data matrix $\mathbf{T}$ we have

$$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{W} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{T}$$

Solving for $\mathbf{W}$ then gives (3.15) as required.

The maximum likelihood solution for $\boldsymbol{\Sigma}$ is easily found by appealing to the standard result from Chapter 2 giving

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))^{\mathrm{T}}.$$

as required. Since we are finding a joint maximum with respect to both $\mathbf{W}$ and $\boldsymbol{\Sigma}$ we see that it is $\mathbf{W}_{\mathrm{ML}}$ which appears in this expression, as in the standard result for an unconditional Gaussian distribution.

3.7 From Bayes' theorem we have

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w}),$$

where the factors on the r.h.s. are given by (3.10) and (3.48), respectively. Writing this out in full, we get

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}) &\propto \left[\prod_{n=1}^{N} \mathcal{N}\left(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right)\right] \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0\right) \\
&\propto \exp\left(-\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})\right) \\
&\quad \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right) \\
&= \exp\Big(-\frac{1}{2}\big(\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}+\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\mathbf{w} - \beta\mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - \beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \beta\mathbf{t}^{\mathrm{T}}\mathbf{t} \\
&\quad \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}+\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\mathbf{w} - \left(\mathbf{S}_0^{-1}\mathbf{m}_0+\beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right)^{\mathrm{T}}\mathbf{w} \\
&\quad -\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}\mathbf{m}_0+\beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right) + \beta\mathbf{t}^{\mathrm{T}}\mathbf{t} + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0\big)\Big) \\
&= \exp\left(-\frac{1}{2}\left(\mathbf{w}-\mathbf{m}_N\right)^{\mathrm{T}}\mathbf{S}_N^{-1}\left(\mathbf{w}-\mathbf{m}_N\right)\right) \\
&\quad \exp\left(-\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\right)\right)
\end{aligned}$$

where we have used (3.50) and (3.51) when completing the square in the last step. The first exponential corrsponds to the posterior, unnormalized Gaussian distribution over $\mathbf{w}$, while the second exponential is independent of $\mathbf{w}$ and hence can be absorbed into the normalization factor.

3.8 Combining the prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

and the likelihood

$$p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left(-\frac{\beta}{2}(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2\right) \tag{130}$$

where $\boldsymbol{\phi}_{N+1} = \boldsymbol{\phi}(\mathbf{x}_{N+1})$, we obtain a posterior of the form

$$\begin{aligned}
&p(\mathbf{w}|t_{N+1}, \mathbf{x}_{N+1}, \mathbf{m}_N, \mathbf{S}_N) \\
&\quad \propto \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N) - \frac{1}{2}\beta(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2\right).
\end{aligned}$$

We can expand the argument of the exponential, omitting the $-1/2$ factors, as follows

$$\begin{aligned}
&(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)+\beta(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2\\
&\quad=\mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w}-2\mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\\
&\qquad+\beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}\mathbf{w}-2\beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}t_{N+1}+\text{const}\\
&\quad=\mathbf{w}^{\mathrm{T}}(\mathbf{S}_N^{-1}+\beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}})\mathbf{w}-2\mathbf{w}^{\mathrm{T}}(\mathbf{S}_N^{-1}\mathbf{m}_N+\beta\boldsymbol{\phi}_{N+1}t_{N+1})+\text{const},
\end{aligned}$$

where const denotes remaining terms independent of $\mathbf{w}$. From this we can read off the desired result directly,

$$p(\mathbf{w}|t_{N+1},\mathbf{x}_{N+1},\mathbf{m}_N,\mathbf{S}_N)=\mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1},\mathbf{S}_{N+1}),$$

with

$$\mathbf{S}_{N+1}^{-1}=\mathbf{S}_N^{-1}+\beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}. \tag{131}$$

and

$$\mathbf{m}_{N+1}=\mathbf{S}_{N+1}(\mathbf{S}_N^{-1}\mathbf{m}_N+\beta\boldsymbol{\phi}_{N+1}t_{N+1}). \tag{132}$$

3.9 Identifying (2.113) with (3.49) and (2.114) with (130), such that

$$\mathbf{x}\Rightarrow\mathbf{w}\quad\boldsymbol{\mu}\Rightarrow\mathbf{m}_N\quad\boldsymbol{\Lambda}^{-1}\Rightarrow\mathbf{S}_N$$

$$\mathbf{y}\Rightarrow t_{N+1}\quad\mathbf{A}\Rightarrow\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}=\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\quad\mathbf{b}\Rightarrow\mathbf{0}\quad\mathbf{L}^{-1}\Rightarrow\beta\mathbf{I},$$

(2.116) and (2.117) directly give

$$p(\mathbf{w}|t_{N+1},\mathbf{x}_{N+1})=\mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1},\mathbf{S}_{N+1})$$

where $\mathbf{S}_{N+1}$ and $\mathbf{m}_{N+1}$ are given by (131) and (132), respectively.

3.10 Using (3.3), (3.8) and (3.49), we can re-write (3.57) as

$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta)=\int\mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w},\beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)\,\mathrm{d}\mathbf{w}.$$

By matching the first factor of the integrand with (2.114) and the second factor with (2.113), we obtain the desired result directly from (2.115).

3.11 From (3.59) we have

$$\sigma_{N+1}^2(\mathbf{x})=\frac{1}{\beta}+\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_{N+1}\boldsymbol{\phi}(\mathbf{x}) \tag{133}$$

where $\mathbf{S}_{N+1}$ is given by (131). From (131) and (3.110) we get

$$\begin{aligned}
\mathbf{S}_{N+1}&=\left(\mathbf{S}_N^{-1}+\beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\right)^{-1}\\
&=\mathbf{S}_N-\frac{\left(\mathbf{S}_N\boldsymbol{\phi}_{N+1}\beta^{1/2}\right)\left(\beta^{1/2}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\right)}{1+\beta\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}\\
&=\mathbf{S}_N-\frac{\beta\mathbf{S}_N\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N}{1+\beta\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}.
\end{aligned}$$

Using this and (3.59), we can rewrite (133) as

$$\begin{aligned}\sigma^2_{N+1}(\mathbf{x}) &= \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\left(\mathbf{S}_N - \frac{\beta\mathbf{S}_N\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}^{\mathrm{T}}_{N+1}\mathbf{S}_N}{1+\beta\boldsymbol{\phi}^{\mathrm{T}}_{N+1}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}\right)\boldsymbol{\phi}(\mathbf{x}) \\ &= \sigma^2_N(\mathbf{x}) - \frac{\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}^{\mathrm{T}}_{N+1}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})}{1+\beta\boldsymbol{\phi}^{\mathrm{T}}_{N+1}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}. \qquad (134)\end{aligned}$$

Since $\mathbf{S}_N$ is positive definite, the numerator and denominator of the second term in (134) will be non-negative and positive, respectively, and hence $\sigma^2_{N+1}(\mathbf{x}) \leqslant \sigma^2_N(\mathbf{x})$.

3.12 It is easiest to work in log space. The log of the posterior distribution is given by

$$\begin{aligned}\ln p(\mathbf{w},\beta|\mathbf{t}) &= \ln p(\mathbf{w},\beta) + \sum_{n=1}^{N}\ln p(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \frac{M}{2}\ln\beta - \frac{1}{2}\ln|\mathbf{S}_0| - \frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\ &\quad -b_0\beta + (a_0-1)\ln\beta \\ &\quad +\frac{N}{2}\ln\beta - \frac{\beta}{2}\sum_{n=1}^{N}\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) - t_n\}^2 + \text{const.}\end{aligned}$$

Using the product rule, the posterior distribution can be written as $p(\mathbf{w},\beta|\mathbf{t}) = p(\mathbf{w}|\beta,\mathbf{t})p(\beta|\mathbf{t})$. Consider first the dependence on $\mathbf{w}$. We have

$$\ln p(\mathbf{w}|\beta,\mathbf{t}) = -\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\left[\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \mathbf{S}_0^{-1}\right]\mathbf{w} + \mathbf{w}^{\mathrm{T}}\left[\beta\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right] + \text{const.}$$

Thus we see that $p(\mathbf{w}|\beta,\mathbf{t})$ is a Gaussian distribution with mean and covariance given by

$$\mathbf{m}_N = \mathbf{S}_N\left[\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right] \qquad (135)$$

$$\beta\mathbf{S}_N^{-1} = \beta\left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right). \qquad (136)$$

To find $p(\beta|\mathbf{t})$ we first need to complete the square over $\mathbf{w}$ to ensure that we pick up all terms involving β (any terms independent of β may be discarded since these will be absorbed into the normalization coefficient which itself will be found by inspection at the end). We also need to remember that a factor of $(M/2)\ln\beta$ will be absorbed by the normalisation factor of $p(\mathbf{w}|\beta,\mathbf{t})$. Thus

$$\begin{aligned}\ln p(\beta|\mathbf{t}) = &-\frac{\beta}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 + \frac{\beta}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N \\ &+\frac{N}{2}\ln\beta - b_0\beta + (a_0-1)\ln\beta - \frac{\beta}{2}\sum_{n=1}^{N}t_n^2 + \text{const.}\end{aligned}$$

We recognize this as the log of a Gamma distribution. Reading off the coefficients of β and $\ln \beta$ we then have

$$a_N = a_0 + \frac{N}{2} \tag{137}$$

$$b_N = b_0 + \frac{1}{2}\left(\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \sum_{n=1}^{N} t_n^2\right). \tag{138}$$

3.13 Following the line of presentation from Section 3.3.2, the predictive distribution is now given by

$$p(t|\mathbf{x},\mathbf{t}) = \iint \mathcal{N}\left(t|\phi(\mathbf{x})^{\mathrm{T}}\mathbf{w},\beta^{-1}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N\right)\,\mathrm{d}\mathbf{w} \\ \mathrm{Gam}\left(\beta|a_N,b_N\right)\,\mathrm{d}\beta \tag{139}$$

We begin by performing the integral over $\mathbf{w}$. Identifying (2.113) with (3.49) and (2.114) with (3.8), using (3.3), such that

$$\mathbf{x} \Rightarrow \mathbf{w} \quad \boldsymbol{\mu} \Rightarrow \mathbf{m}_N \quad \boldsymbol{\Lambda}^{-1} \Rightarrow \mathbf{S}_N$$

$$\mathbf{y} \Rightarrow t \quad \mathbf{A} \Rightarrow \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} = \boldsymbol{\phi}^{\mathrm{T}} \quad \mathbf{b} \Rightarrow 0 \quad \mathbf{L}^{-1} \Rightarrow \beta^{-1},$$

(2.115) and (136) give

$$\begin{aligned} p(t|\beta) &= \mathcal{N}\left(t|\boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N,\beta^{-1} + \boldsymbol{\phi}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}\right) \\ &= \mathcal{N}\left(t|\boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N,\beta^{-1}\left(1 + \boldsymbol{\phi}^{\mathrm{T}}(\mathbf{S}_0 + \boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}\right)\right). \end{aligned}$$

Substituting this back into (139) we get

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t}) = \int \mathcal{N}\left(t|\boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N,\beta^{-1}s\right)\mathrm{Gam}\left(\beta|a_N,b_N\right)\,\mathrm{d}\beta,$$

where we have defined

$$s = 1 + \boldsymbol{\phi}^{\mathrm{T}}(\mathbf{S}_0 + \boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}.$$

We can now use (2.158)– (2.160) to obtain the final result:

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t}) = \mathrm{St}\left(t|\mu,\lambda,\nu\right)$$

where

$$\mu = \boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N \quad \lambda = \frac{a_N}{b_N}s^{-1} \quad \nu = 2a_N.$$

3.14 For $\alpha = 0$ the covariance matrix $\mathbf{S}_N$ becomes

$$\mathbf{S}_N = (\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}. \tag{140}$$

Let us define a new set of orthonormal basis functions given by linear combinations of the original basis functions so that

$$\boldsymbol{\psi}(\mathbf{x}) = \mathbf{V}\boldsymbol{\phi}(\mathbf{x}) \tag{141}$$

where $\mathbf{V}$ is an $M \times M$ matrix. Since both the original and the new basis functions are linearly independent and span the same space, this matrix must be invertible and hence

$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{V}^{-1}\boldsymbol{\psi}(\mathbf{x}).$$

For the data set $\{\mathbf{x}_n\}$, (141) and (3.16) give

$$\boldsymbol{\Psi} = \boldsymbol{\Phi}\mathbf{V}^{\mathrm{T}}$$

and consequently

$$\boldsymbol{\Phi} = \boldsymbol{\Psi}\mathbf{V}^{-\mathrm{T}}$$

where $\mathbf{V}^{-\mathrm{T}}$ denotes $(\mathbf{V}^{-1})^{\mathrm{T}}$. Orthonormality implies

$$\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} = \mathbf{I}.$$

Note that $(\mathbf{V}^{-1})^{\mathrm{T}} = (\mathbf{V}^{\mathrm{T}})^{-1}$ as is easily verified. From (140), the covariance matrix then becomes

$$\mathbf{S}_N = \beta^{-1}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1} = \beta^{-1}(\mathbf{V}^{-\mathrm{T}}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\mathbf{V}^{-1})^{-1} = \beta^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V}.$$

Here we have used the orthonormality of the $\psi_i(\mathbf{x})$. Hence the equivalent kernel becomes

$$k(\mathbf{x}, \mathbf{x}') = \beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{V}\boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}')$$

as required. From the orthonormality condition, and setting $j = 1$, it follows that

$$\sum_{n=1}^{N} \psi_i(\mathbf{x}_n)\psi_1(\mathbf{x}_n) = \sum_{n=1}^{N} \psi_i(\mathbf{x}_n) = \delta_{i1}$$

where we have used $\psi_1(\mathbf{x}) = 1$. Now consider the sum

$$\begin{aligned}\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) &= \sum_{n=1}^{N} \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_n) = \sum_{n=1}^{N}\sum_{i=1}^{M} \psi_i(\mathbf{x})\psi_i(\mathbf{x}_n) \\ &= \sum_{i=1}^{M} \psi_i(\mathbf{x})\delta_{i1} = \psi_1(\mathbf{x}) = 1\end{aligned}$$

which proves the summation constraint as required.

3.15 This is easily shown by substituting the re-estimation formulae (3.92) and (3.95) into (3.82), giving

$$\begin{aligned} E(\mathbf{m}_N) &= \frac{\beta}{2}\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N \\ &= \frac{N-\gamma}{2} + \frac{\gamma}{2} = \frac{N}{2}\,. \end{aligned}$$

3.16 The likelihood function is a product of independent univariate Gaussians and so can be written as a joint Gaussian distribution over $\mathbf{t}$ with diagonal covariance matrix in the form

$$p(\mathbf{t}|\mathbf{w},\beta) = \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}_N). \tag{142}$$

Identifying (2.113) with the prior distribution $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I})$ and (2.114) with (142), such that

$$\mathbf{x} \Rightarrow \mathbf{w} \quad \boldsymbol{\mu} \Rightarrow \mathbf{0} \quad \boldsymbol{\Lambda}^{-1} \Rightarrow \alpha^{-1}\mathbf{I}_M$$

$$\mathbf{y} \Rightarrow \mathbf{t} \quad \mathbf{A} \Rightarrow \boldsymbol{\Phi} \quad \mathbf{b} \Rightarrow \mathbf{0} \quad \mathbf{L}^{-1} \Rightarrow \beta^{-1}\mathbf{I}_N,$$

(2.115) gives

$$p(\mathbf{t}|\alpha,\beta) = \mathcal{N}(\mathbf{t}|\mathbf{0},\beta^{-1}\mathbf{I}_N + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}).$$

Taking the log we obtain

$$\begin{aligned} \ln p(\mathbf{t}|\alpha,\beta) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\beta^{-1}\mathbf{I}_N + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\right| \\ - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta^{-1}\mathbf{I}_N + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\right)\mathbf{t}. \end{aligned} \tag{143}$$

Using the result (C.14) for the determinant we have

$$\begin{aligned} \left|\beta^{-1}\mathbf{I}_N + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\right| &= \beta^{-N}\left|\mathbf{I}_N + \beta\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\right| \\ &= \beta^{-N}\left|\mathbf{I}_M + \beta\alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right| \\ &= \beta^{-N}\alpha^{-M}\left|\alpha\mathbf{I}_M + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right| \\ &= \beta^{-N}\alpha^{-M}\left|\mathbf{A}\right| \end{aligned}$$

where we have used (3.81). Next consider the quadratic term in $\mathbf{t}$ and make use of the identity (C.7) together with (3.81) and (3.84) to give

$$\begin{aligned} &-\frac{1}{2}\mathbf{t}\left(\beta^{-1}\mathbf{I}_N + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\right)^{-1}\mathbf{t} \\ &= -\frac{1}{2}\mathbf{t}^{\mathrm{T}}\left[\beta\mathbf{I}_N - \beta\boldsymbol{\Phi}\left(\alpha\mathbf{I}_M + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\beta\right]\mathbf{t} \\ &= -\frac{\beta}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{\beta^2}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \\ &= -\frac{\beta}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N \\ &= -\frac{\beta}{2}\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}_N\|^2 - \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N \end{aligned}$$

where in the last step, we have exploited results from Solution 3.18. Substituting for the determinant and the quadratic term in (143) we obtain (3.86).

3.17 Using (3.11), (3.12) and (3.52) together with the definition for the Gaussian, (2.43), we can rewrite (3.77) as follows:

$$\begin{aligned} p(\mathbf{t}|\alpha,\beta) &= \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w} \\ &= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left(-\beta E_D(\mathbf{w})\right)\exp\left(-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)\,\mathrm{d}\mathbf{w} \\ &= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left(-E(\mathbf{w})\right)\,\mathrm{d}\mathbf{w}, \end{aligned}$$

where $E(\mathbf{w})$ is defined by (3.79).

3.18 We can rewrite (3.79)

$$\begin{aligned} &\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi}\mathbf{w}\|^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &\quad= \frac{\beta}{2}\left(\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w}+\mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w}\right)+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &\quad= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w}+\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right) \end{aligned}$$

where, in the last line, we have used (3.81). We now use the tricks of adding $\mathbf{0} = \mathbf{m}_{\mathbf{N}}^{\mathrm{T}}\mathbf{A}\mathbf{m}_{\mathbf{N}} - \mathbf{m}_{\mathbf{N}}^{\mathrm{T}}\mathbf{A}\mathbf{m}_{\mathbf{N}}$ and using $\mathbf{I} = \mathbf{A}^{-1}\mathbf{A}$, combined with (3.84), as follows:

$$\begin{aligned} &\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w}+\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right) \\ &\quad= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{A}\mathbf{w}+\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right) \\ &\quad= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{w}+\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}+\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right) \\ &\quad= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)+\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N). \end{aligned}$$

Here the last term equals term the last term of (3.80) and so it remains to show that the first term equals the r.h.s. of (3.82). To do this, we use the same tricks again:

$$\begin{aligned} \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right) &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N+\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right) \\ &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{A}^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\beta+\mathbf{m}_N^{\mathrm{T}}\left(\alpha\mathbf{I}+\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)\mathbf{m}_N\right) \\ &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\beta+\beta\mathbf{m}_N^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{m}_N+\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right) \\ &= \frac{1}{2}\left(\beta(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N)^{\mathrm{T}}(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N)+\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right) \\ &= \frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N\|^2+\frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N \end{aligned}$$

as required.

3.19 From (3.80) we see that the integrand of (3.85) is an unnormalized Gaussian and hence integrates to the inverse of the corresponding normalizing constant, which can be read off from the r.h.s. of (2.43) as

$$(2\pi)^{M/2}\,|\mathbf{A}^{-1}|^{1/2}.$$

Using (3.78), (3.85) and the properties of the logarithm, we get

$$\begin{aligned}\ln p(\mathbf{t}|\alpha,\beta) &= \frac{M}{2}(\ln\alpha - \ln(2\pi)) + \frac{N}{2}(\ln\beta - \ln(2\pi)) + \ln\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w} \\ &= \frac{M}{2}(\ln\alpha - \ln(2\pi)) + \frac{N}{2}(\ln\beta - \ln(2\pi)) - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| + \frac{M}{2}\ln(2\pi)\end{aligned}$$

which equals (3.86).

3.20 We only need to consider the terms of (3.86) that depend on α, which are the first, third and fourth terms.

Following the sequence of steps in Section 3.5.2, we start with the last of these terms,

$$-\frac{1}{2}\ln|\mathbf{A}|.$$

From (3.81), (3.87) and the fact that that eigenvectors $\mathbf{u}_i$ are orthonormal (see also Appendix C), we find that the eigenvectors of $\mathbf{A}$ to be $\alpha+\lambda_i$. We can then use (C.47) and the properties of the logarithm to take us from the left to the right side of (3.88).

The derivatives for the first and third term of (3.86) are more easily obtained using standard derivatives and (3.82), yielding

$$\frac{1}{2}\left(\frac{M}{\alpha} + \mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right).$$

We combine these results into (3.89), from which we get (3.92) via (3.90). The expression for γ in (3.91) is obtained from (3.90) by substituting

$$\sum_i^M \frac{\lambda_i + \alpha}{\lambda_i + \alpha}$$

for M and re-arranging.

3.21 The eigenvector equation for the $M \times M$ real, symmetric matrix $\mathbf{A}$ can be written as

$$\mathbf{A}\mathbf{u}_i = \eta_i\mathbf{u}_i$$

where $\{\mathbf{u}_i\}$ are a set of M orthonormal vectors, and the M eigenvalues $\{\eta_i\}$ are all real. We first express the left hand side of (3.117) in terms of the eigenvalues of $\mathbf{A}$. The log of the determinant of $\mathbf{A}$ can be written as

$$\ln|\mathbf{A}| = \ln\prod_{i=1}^{M}\eta_i = \sum_{i=1}^{M}\ln\eta_i.$$

Taking the derivative with respect to some scalar α we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln|\mathbf{A}| = \sum_{i=1}^{M}\frac{1}{\eta_i}\frac{\mathrm{d}}{\mathrm{d}\alpha}\eta_i.$$

We now express the right hand side of (3.117) in terms of the eigenvector expansion and show that it takes the same form. First we note that $\mathbf{A}$ can be expanded in terms of its own eigenvectors to give

$$\mathbf{A} = \sum_{i=1}^{M}\eta_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$$

and similarly the inverse can be written as

$$\mathbf{A}^{-1} = \sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}.$$

Thus we have

$$\begin{aligned}
\mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right) &= \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{j=1}^{M}\eta_j\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right) \\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\left\{\sum_{j=1}^{M}\frac{\mathrm{d}\eta_j}{\mathrm{d}\alpha}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}} + \eta_j\left(\mathbf{b}_j\mathbf{u}_j^{\mathrm{T}} + \mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\right\}\right) \\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\sum_{j=1}^{M}\frac{\mathrm{d}\eta_j}{\mathrm{d}\alpha}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right) \\
&\quad + \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\sum_{j=1}^{M}\eta_j\left(\mathbf{b}_j\mathbf{u}_j^{\mathrm{T}} + \mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\right) \qquad (144)
\end{aligned}$$

where $\mathbf{b}_j = \mathrm{d}\mathbf{u}_j/\mathrm{d}\alpha$. Using the properties of the trace and the orthognality of

eigenvectors, we can rewrite the second term as

$$\begin{aligned}
&\mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\sum_{j=1}^{M}\eta_j\left(\mathbf{b}_j\mathbf{u}_j^{\mathrm{T}}+\mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\right)\\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\sum_{j=1}^{M}2\eta_j\mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{2\eta_j}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\left(\mathbf{b}_j\mathbf{u}_j^{\mathrm{T}}+\mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\right)\\
&= \mathrm{Tr}\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{i}^{M}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\right).
\end{aligned}$$

However,

$$\sum_{i}^{M}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}=\mathbf{I}$$

which is constant and thus its derivative w.r.t. α will be zero and the second term in (144) vanishes.

For the first term in (144), we again use the properties of the trace and the orthognality of eigenvectors to obtain

$$\mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)=\sum_{i=1}^{M}\frac{1}{\eta_i}\frac{\mathrm{d}\eta_i}{\mathrm{d}\alpha}.$$

We have now shown that both the left and right hand sides of (3.117) take the same form when expressed in terms of the eigenvector expansion. Next, we use (3.117) to differentiate (3.86) w.r.t. α, yielding

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln p(\mathbf{t}|\alpha\beta) &= \frac{M}{2}\frac{1}{\alpha}-\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N-\frac{1}{2}\mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)\\
&= \frac{1}{2}\left(\frac{M}{\alpha}-\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N-\mathrm{Tr}\left(\mathbf{A}^{-1}\right)\right)\\
&= \frac{1}{2}\left(\frac{M}{\alpha}-\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N-\sum_{i}\frac{1}{\lambda_i+\alpha}\right)
\end{aligned}$$

which we recognize as the r.h.s. of (3.89), from which (3.92) can be derived as detailed in Section 3.5.2, immediately following (3.89).

3.22 Using (3.82) and (3.93)—the derivation of latter is detailed in Section 3.5.2—we get the derivative of (3.86) w.r.t. β as the r.h.s. of (3.94). Rearranging this, collecting the β-dependent terms on one side of the equation and the remaining term on the other, we obtain (3.95).

3.23 From (3.10), (3.112) and the properties of the Gaussian and Gamma distributions (see Appendix B), we get

$$\begin{aligned}
p(\mathbf{t}) &= \iint p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\beta)\,\mathrm{d}\mathbf{w}p(\beta)\,\mathrm{d}\beta \\
&= \iint \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left\{-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})\right\} \\
&\qquad \left(\frac{\beta}{2\pi}\right)^{M/2} |\mathbf{S}_0|^{-1/2} \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right\}\,\mathrm{d}\mathbf{w} \\
&\qquad \Gamma(a_0)^{-1} b_0^{a_0} \beta^{a_0-1} \exp(-b_0\beta)\,\mathrm{d}\beta \\
&= \frac{b_0^{a_0}}{\left((2\pi)^{M+N}|\mathbf{S}_0|\right)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})\right\} \\
&\qquad \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right\}\,\mathrm{d}\mathbf{w} \\
&\qquad \beta^{a_0-1}\beta^{N/2}\beta^{M/2}\exp(-b_0\beta)\,\mathrm{d}\beta \\
&= \frac{b_0^{a_0}}{\left((2\pi)^{M+N}|\mathbf{S}_0|\right)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right\}\,\mathrm{d}\mathbf{w} \\
&\qquad \exp\left\{-\frac{\beta}{2}\left(\mathbf{t}^{\mathrm{T}}\mathbf{t}+\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0-\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\right)\right\} \\
&\qquad \beta^{a_N-1}\beta^{M/2}\exp(-b_0\beta)\,\mathrm{d}\beta
\end{aligned}$$

where we have completed the square for the quadratic form in $\mathbf{w}$, using

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left[\mathbf{S}_0^{-1}\mathbf{m}_0+\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\right] \\
\mathbf{S}_N^{-1} &= \beta\left(\mathbf{S}_0^{-1}+\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right) \\
a_N &= a_0+\frac{N}{2} \\
b_N &= b_0+\frac{1}{2}\left(\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0-\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N+\sum_{n=1}^{N}t_n^2\right).
\end{aligned}$$

Now we are ready to do the integration, first over $\mathbf{w}$ and then β, and re-arrange the

terms to obtain the desired result

$$\begin{aligned} p(\mathbf{t}) &= \frac{b_0^{a_0}}{\left((2\pi)^{M+N}|\mathbf{S}_0|\right)^{1/2}}(2\pi)^{M/2}|\mathbf{S}_N|^{1/2}\int \beta^{a_N-1}\exp(-b_N\beta)\,\mathrm{d}\beta \\ &= \frac{1}{(2\pi)^{N/2}}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{\Gamma(a_N)}{\Gamma(a_0)}. \end{aligned}$$

3.24 Substituting the r.h.s. of (3.10), (3.112) and (3.113) into (3.119), we get

$$p(\mathbf{t}) = \frac{\mathcal{N}\left(\mathbf{t}|\mathbf{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{m}_0,\beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}\left(\beta|a_0,b_0\right)}{\mathcal{N}\left(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}\left(\beta|a_N,b_N\right)}. \tag{145}$$

Using the definitions of the Gaussian and Gamma distributions, we can write this as

$$\begin{aligned} &\left(\frac{\beta}{2\pi}\right)^{N/2}\exp\left(-\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi}\mathbf{w}\|^2\right) \\ &\qquad\left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_0|^{1/2}\exp\left(-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right) \\ &\qquad\qquad\Gamma(a_0)^{-1}b_0^{a_0}\beta^{a_0-1}\exp(-b_0\beta) \\ &\left\{\left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\left(-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right)\right. \\ &\qquad\qquad\left.\Gamma(a_N)^{-1}b_N^{a_N}\beta^{a_N-1}\exp(-b_N\beta)\right\}^{-1}. \end{aligned} \tag{146}$$

Concentrating on the factors corresponding to the denominator in (145), i.e. the fac-

tors inside $\{\ldots\})^{-1}$ in (146), we can use (135)–(138) to get

$$\begin{aligned}
&\mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}\left(\beta|a_N, b_N\right)\\
&= \left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\Bigg(-\frac{\beta}{2}\big(\mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w}\\
&\quad + \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\big)\Bigg)\Gamma(a_N)^{-1}b_N^{a_N}\beta^{a_N-1}\exp(-b_N\beta)\\
&= \left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\Bigg(-\frac{\beta}{2}\big(\mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0\\
&\quad - \mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} - \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w} - \mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} + \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\big)\Bigg)\\
&\quad \Gamma(a_N)^{-1}b_N^{a_N}\beta^{a_0+N/2-1}\\
&\quad \exp\left(-\left(b_0 + \frac{1}{2}\left(\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \mathbf{t}^{\mathrm{T}}\mathbf{t}\right)\right)\beta\right)\\
&= \left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\left(-\frac{\beta}{2}\left((\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0(\mathbf{w}-\mathbf{m}_0) + \|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\|^2\right)\right)\\
&\quad \Gamma(a_N)^{-1}b_N^{a_N}\beta^{a_N+N/2-1}\exp(-b_0\beta).
\end{aligned}$$

Substituting this into (146), the exponential factors along with $\beta^{a_0+N/2-1}(\beta/2\pi)^{M/2}$ cancel and we are left with (3.118).

Chapter 4 Linear Models for Classification

4.1 Assume that the convex hulls of $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ intersect. Then there exist a point $\mathbf{z}$ such that

$$\mathbf{z} = \sum_n \alpha_n\mathbf{x}_n = \sum_m \beta_m\mathbf{y}_m$$

where $\beta_m \geqslant 0$ for all m and $\sum_m \beta_m = 1$. If $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ also were to be linearly separable, we would have that

$$\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{z} + w_0 = \sum_n \alpha_n\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}_n + w_0 = \sum_n \alpha_n($$

since $\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}_n + w_0 > 0$ and the $\{\alpha_n\}$ are all non-negative and sum to 1, but by the corresponding argument

$$\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{z} + w_0 = \sum_m \beta_m\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0 = \sum_m \beta_m(\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0) < 0,$$

which is a contradiction and hence $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ cannot be linearly separable if their convex hulls intersect.

If we instead assume that $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ are linearly separable and consider a point $\mathbf{z}$ in the intersection of their convex hulls, the same contradiction arise. Thus no such point can exist and the intersection of the convex hulls of $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ must be empty.

4.2 For the purpose of this exercise, we make the contribution of the bias weights explicit in (4.15), giving

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})\right\}, \tag{147}$$

where $\mathbf{w}_0$ is the column vector of bias weights (the top row of $\widetilde{\mathbf{W}}$ transposed) and $\mathbf{1}$ is a column vector of N ones.

We can take the derivative of (147) w.r.t. $\mathbf{w}_0$, giving

$$2N\mathbf{w}_0 + 2(\mathbf{X}\mathbf{W} - \mathbf{T})^{\mathrm{T}}\mathbf{1}.$$

Setting this to zero, and solving for $\mathbf{w}_0$, we obtain

$$\mathbf{w}_0 = \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \tag{148}$$

where

$$\bar{\mathbf{t}} = \frac{1}{N}\mathbf{T}^{\mathrm{T}}\mathbf{1} \quad \text{and} \quad \bar{\mathbf{x}} = \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{1}.$$

If we subsitute (148) into (147), we get

$$E_D(\mathbf{W}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{X}\mathbf{W} + \overline{\mathbf{T}} - \overline{\mathbf{X}}\mathbf{W} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \overline{\mathbf{T}} - \overline{\mathbf{X}}\mathbf{W} - \mathbf{T})\right\},$$

where

$$\overline{\mathbf{T}} = \mathbf{1}\bar{\mathbf{t}}^{\mathrm{T}} \quad \text{and} \quad \overline{\mathbf{X}} = \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}.$$

Setting the derivative of this w.r.t. $\mathbf{W}$ to zero we get

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

where we have defined $\widehat{\mathbf{X}} = \mathbf{X} - \overline{\mathbf{X}}$ and $\widehat{\mathbf{T}} = \mathbf{T} - \overline{\mathbf{T}}$.

Now consider the prediction for a new input vector $\mathbf{x}^\star$,

$$\begin{aligned}
\mathbf{y}(\mathbf{x}^\star) &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \mathbf{w}_0 \\
&= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \\
&= \bar{\mathbf{t}} - \widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}(\mathbf{x}^\star - \bar{\mathbf{x}}).
\end{aligned} \tag{149}$$

If we apply (4.157) to $\bar{\mathbf{t}}$, we get

$$\mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} = \frac{1}{N}\mathbf{a}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{1} = -b.$$

Therefore, applying (4.157) to (149), we obtain

$$\begin{aligned}\mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}^\star) &= \mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} + \mathbf{a}^{\mathrm{T}}\widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^\dagger\right)^{\mathrm{T}}(\mathbf{x}^\star - \bar{\mathbf{x}}) \\ &= \mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} = -b,\end{aligned}$$

since $\mathbf{a}^{\mathrm{T}}\widehat{\mathbf{T}}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}(\mathbf{T} - \overline{\mathbf{T}})^{\mathrm{T}} = b(\mathbf{1} - \mathbf{1})^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$.

4.3 When we consider several simultaneous constraints, (4.157) becomes

$$\mathbf{A}\mathbf{t}_n + \mathbf{b} = \mathbf{0}, \tag{150}$$

where $\mathbf{A}$ is a matrix and $\mathbf{b}$ is a column vector such that each row of $\mathbf{A}$ and element of $\mathbf{b}$ correspond to one linear constraint.

If we apply (150) to (149), we obtain

$$\begin{aligned}\mathbf{A}\mathbf{y}(\mathbf{x}^\star) &= \mathbf{A}\bar{\mathbf{t}} - \mathbf{A}\widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^\dagger\right)^{\mathrm{T}}(\mathbf{x}^\star - \bar{\mathbf{x}}) \\ &= \mathbf{A}\bar{\mathbf{t}} = -\mathbf{b},\end{aligned}$$

since $\mathbf{A}\widehat{\mathbf{T}}^{\mathrm{T}} = \mathbf{A}(\mathbf{T} - \overline{\mathbf{T}})^{\mathrm{T}} = \mathbf{b}\mathbf{1}^{\mathrm{T}} - \mathbf{b}\mathbf{1}^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$. Thus $\mathbf{A}\mathbf{y}(\mathbf{x}^\star) + \mathbf{b} = \mathbf{0}$.

4.4 **NOTE**: In the 1st printing of PRML, the text of the exercise refers equation (4.23) where it should refer to (4.22).

From (4.22) we can construct the Lagrangian function

$$L = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1) + \lambda\left(\mathbf{w}^{\mathrm{T}}\mathbf{w} - 1\right).$$

Taking the gradient of L we obtain

$$\nabla L = \mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} \tag{151}$$

and setting this gradient to zero gives

$$\mathbf{w} = -\frac{1}{2\lambda}(\mathbf{m}_2 - \mathbf{m}_1)$$

form which it follows that $\mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1$.

4.5 Starting with the numerator on the r.h.s. of (4.25), we can use (4.23) and (4.27) to rewrite it as follows:

$$\begin{aligned}(m_2 - m_1)^2 &= \left(\mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)\right)^2 \\ &= \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}\mathbf{w} \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}.\end{aligned} \tag{152}$$

Similarly, we can use (4.20), (4.23), (4.24), and (4.28) to rewrite the denominator of the r.h.s. of (4.25):

$$\begin{aligned} s_1^2 + s_2^2 &= \sum_{n\in\mathcal{C}_1} (y_n - m_1)^2 + \sum_{k\in\mathcal{C}_2} (y_k - m_2)^2 \\ &= \sum_{n\in\mathcal{C}_1} \left(\mathbf{w}^{\mathrm{T}}(\mathbf{x}_n - \mathbf{m}_1)\right)^2 + \sum_{k\in\mathcal{C}_2} \left(\mathbf{w}^{\mathrm{T}}(\mathbf{x}_k - \mathbf{m}_2)\right)^2 \\ &= \sum_{n\in\mathcal{C}_1} \mathbf{w}^{\mathrm{T}}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}}\mathbf{w} \\ &\qquad + \sum_{k\in\mathcal{C}_2} \mathbf{w}^{\mathrm{T}}(\mathbf{x}_k - \mathbf{m}_2)(\mathbf{x}_k - \mathbf{m}_2)^{\mathrm{T}}\mathbf{w} \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}. \end{aligned} \tag{153}$$

Substituting (152) and (153) in (4.25) we obtain (4.26).

4.6 Using (4.21) and (4.34) along with the chosen target coding scheme, we can re-write the l.h.s. of (4.33) as follows:

$$\begin{aligned} \sum_{n=1}^{N} \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - w_0 - t_n\right)\mathbf{x}_n &= \sum_{n=1}^{N} \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - \mathbf{w}^{\mathrm{T}}\mathbf{m} - t_n\right)\mathbf{x}_n \\ &= \sum_{n=1}^{N} \left\{\left(\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \mathbf{x}_n\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} - \mathbf{x}_n t_n\right\} \\ &= \sum_{n\in\mathcal{C}_1} \left\{\left(\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \mathbf{x}_n\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} - \mathbf{x}_n t_n\right\} \\ &\qquad \sum_{m\in\mathcal{C}_2} \left\{\left(\mathbf{x}_m\mathbf{x}_m^{\mathrm{T}} - \mathbf{x}_m\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} - \mathbf{x}_m t_m\right\} \\ &= \left(\sum_{n\in\mathcal{C}_1} \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - N_1\mathbf{m}_1\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} - N_1\mathbf{m}_1\frac{N}{N_1} \\ &\qquad \left(\sum_{m\in\mathcal{C}_2} \mathbf{x}_m\mathbf{x}_m^{\mathrm{T}} - N_2\mathbf{m}_2\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} + N_2\mathbf{m}_2\frac{N}{N_2} \\ &= \left(\sum_{n\in\mathcal{C}_1} \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} + \sum_{m\in\mathcal{C}_2} \mathbf{x}_m\mathbf{x}_m^{\mathrm{T}} - (N_1\mathbf{m}_1 + N_2\mathbf{m}_2)\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} \\ &\qquad -N(\mathbf{m}_1 - \mathbf{m}_2). \end{aligned} \tag{154}$$

We then use the identity

$$\begin{aligned}\sum_{i\in\mathcal{C}_k}(\mathbf{x}_i-\mathbf{m}_k)(\mathbf{x}_i-\mathbf{m}_k)^{\mathrm{T}} &= \sum_{i\in\mathcal{C}_k}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}-\mathbf{x}_i\mathbf{m}_k^{\mathrm{T}}-\mathbf{m}_k\mathbf{x}_i^{\mathrm{T}}+\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}}\right)\\ &= \sum_{i\in\mathcal{C}_k}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}-N_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}}\end{aligned}$$

together with (4.28) and (4.36) to rewrite (154) as

$$\begin{aligned}&\Big(\mathbf{S}_{\mathrm{W}}+N_1\mathbf{m}_1\mathbf{m}_1^{\mathrm{T}}+N_2\mathbf{m}_2\mathbf{m}_2^{\mathrm{T}}\\ &\qquad -(N_1\mathbf{m}_1+N_2\mathbf{m}_2)\frac{1}{N}(N_1\mathbf{m}_1+N_2\mathbf{m}_2)\Big)\mathbf{w}-N(\mathbf{m}_1-\mathbf{m}_2)\\ =\;&\left(\mathbf{S}_{\mathrm{W}}+\left(N_1-\frac{N_1^2}{N}\right)\mathbf{m}_1\mathbf{m}_1^{\mathrm{T}}-\frac{N_1N_2}{N}(\mathbf{m}_1\mathbf{m}_2^{\mathrm{T}}+\mathbf{m}_2\mathbf{m}_1)\right.\\ &\qquad \left.+\left(N_2-\frac{N_2^2}{N}\right)\mathbf{m}_2\mathbf{m}_2^{\mathrm{T}}\right)\mathbf{w}-N(\mathbf{m}_1-\mathbf{m}_2)\\ =\;&\left(\mathbf{S}_{\mathrm{W}}+\frac{(N_1+N_2)N_1-N_1^2}{N}\mathbf{m}_1\mathbf{m}_1^{\mathrm{T}}-\frac{N_1N_2}{N}(\mathbf{m}_1\mathbf{m}_2^{\mathrm{T}}+\mathbf{m}_2\mathbf{m}_1)\right.\\ &\qquad \left.+\frac{(N_1+N_2)N_2-N_2^2}{N}\mathbf{m}_2\mathbf{m}_2^{\mathrm{T}}\right)\mathbf{w}-N(\mathbf{m}_1-\mathbf{m}_2)\\ =\;&\left(\mathbf{S}_{\mathrm{W}}+\frac{N_2N_1}{N}\left(\mathbf{m}_1\mathbf{m}_1^{\mathrm{T}}-\mathbf{m}_1\mathbf{m}_2^{\mathrm{T}}-\mathbf{m}_2\mathbf{m}_1+\mathbf{m}_2\mathbf{m}_2^{\mathrm{T}}\right)\right)\mathbf{w}\\ &\qquad -N(\mathbf{m}_1-\mathbf{m}_2)\\ =\;&\left(\mathbf{S}_{\mathrm{W}}+\frac{N_2N_1}{N}\mathbf{S}_{\mathrm{B}}\right)\mathbf{w}-N(\mathbf{m}_1-\mathbf{m}_2),\end{aligned}$$

where in the last line we also made use of (4.27). From (4.33), this must equal zero, and hence we obtain (4.37).

4.7 From (4.59) we have

$$\begin{aligned}1-\sigma(a) &= 1-\frac{1}{1+e^{-a}}=\frac{1+e^{-a}-1}{1+e^{-a}}\\ &= \frac{e^{-a}}{1+e^{-a}}=\frac{1}{e^{a}+1}=\sigma(-a).\end{aligned}$$

The inverse of the logistic sigmoid is easily found as follows

$$\begin{aligned} y = \sigma(a) &= \frac{1}{1+e^{-a}} \\ \Rightarrow \quad \frac{1}{y} - 1 &= e^{-a} \\ \Rightarrow \quad \ln\left\{\frac{1-y}{y}\right\} &= -a \\ \Rightarrow \quad \ln\left\{\frac{y}{1-y}\right\} &= a = \sigma^{-1}(y). \end{aligned}$$

4.8 Substituting (4.64) into (4.58), we see that the normalizing constants cancel and we are left with

$$\begin{aligned} a &= \ln \frac{\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)\right)p(\mathcal{C}_1)}{\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_2\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_2\right)\right)p(\mathcal{C}_2)} \\ &= -\frac{1}{2}\left(\mathbf{x}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{x} - \mathbf{x}\boldsymbol{\Sigma}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x} + \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}_1\right. \\ &\qquad \left. -\mathbf{x}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{x} + \mathbf{x}\boldsymbol{\Sigma}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x} - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}_2\right) + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\ &= \left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\left(\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}_2\right) + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}. \end{aligned}$$

Substituting this into the rightmost form of (4.57) we obtain (4.65), with $\mathbf{w}$ and w_0 given by (4.66) and (4.67), respectively.

4.9 The likelihood function is given by

$$p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left\{p(\phi_n|\mathcal{C}_k)\pi_k\right\}^{t_{nk}}$$

and taking the logarithm, we obtain

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln p(\phi_n|\mathcal{C}_k) + \ln \pi_k\right\}. \tag{155}$$

In order to maximize the log likelihood with respect to π_k we need to preserve the constraint $\sum_k \pi_k = 1$. This can be done by introducing a Lagrange multiplier λ and maximizing

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).$$

Setting the derivative with respect to π_k equal to zero, we obtain

$$\sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda = 0.$$

Re-arranging then gives

$$-\pi_k \lambda = \sum_{n=1}^{N} t_{nk} = N_k. \tag{156}$$

Summing both sides over k we find that $\lambda = -N$, and using this to eliminate λ we obtain (4.159).

4.10 If we substitute (4.160) into (155) and then use the definition of the multivariate Gaussian, (2.43), we obtain

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \\ -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln|\boldsymbol{\Sigma}| + (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu})\right\}, \tag{157}$$

where we have dropped terms independent of $\{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\Sigma}$.

Setting the derivative of the r.h.s. of (157) w.r.t. $\boldsymbol{\mu}_k$, obtained by using (C.19), to zero, we get

$$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) = 0.$$

Making use of (156), we can re-arrange this to obtain (4.161).

Rewriting the r.h.s. of (157) as

$$-\frac{1}{2}b\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln|\boldsymbol{\Sigma}| + \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi} - \boldsymbol{\mu}_k)^{\mathrm{T}}\right]\right\},$$

we can use (C.24) and (C.28) to calculate the derivative w.r.t. $\boldsymbol{\Sigma}^{-1}$. Setting this to zero we obtain

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{k}^{\mathrm{T}} t_{nk}\left\{\boldsymbol{\Sigma} - (\boldsymbol{\phi}_n - \boldsymbol{\mu}_n)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\right\} = 0.$$

Again making use of (156), we can re-arrange this to obtain (4.162), with $\mathbf{S}_k$ given by (4.163).

Note that, as in Exercise 2.34, we do not enforce that $\boldsymbol{\Sigma}$ should be symmetric, but simply note that the solution is automatically symmetric.

4.11 The generative model for $\boldsymbol{\phi}$ corresponding to the chosen coding scheme is given by

$$p(\boldsymbol{\phi} \mid \mathcal{C}_k) = \prod_{m=1}^{M} p(\boldsymbol{\phi}_m \mid \mathcal{C}_k)$$

where

$$p(\boldsymbol{\phi}_m \mid \mathcal{C}_k) = \prod_{l=1}^{L} \mu_{kml}^{\phi_{ml}},$$

where in turn $\{\mu_{kml}\}$ are the parameters of the multinomial models for $\boldsymbol{\phi}$. Substituting this into (4.63) we see that

$$\begin{aligned} a_k &= \ln p(\boldsymbol{\phi} \mid \mathcal{C}_k)\, p(\mathcal{C}_k) \\ &= \ln p(\mathcal{C}_k) + \sum_{m=1}^{M} \ln p(\boldsymbol{\phi}_m \mid \mathcal{C}_k) \\ &= \ln p(\mathcal{C}_k) + \sum_{m=1}^{M} \sum_{l=1}^{L} \phi_{ml} \ln \mu_{kml}, \end{aligned}$$

which is linear in ϕ_{ml}.

4.12 Differentiating (4.59) we obtain

$$\begin{aligned} \frac{d\sigma}{da} &= \frac{e^{-a}}{(1+e^{-a})^2} \\ &= \sigma(a)\left\{\frac{e^{-a}}{1+e^{-a}}\right\} \\ &= \sigma(a)\left\{\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right\} \\ &= \sigma(a)(1-\sigma(a)). \end{aligned}$$

4.13 We start by computing the derivative of (4.90) w.r.t. y_n

$$\begin{aligned} \frac{\partial E}{\partial y_n} &= \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} && (158) \\ &= \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)} \\ &= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} && (159) \\ &= \frac{y_n - t_n}{y_n(1-y_n)}. && (160) \end{aligned}$$

From (4.88), we see that

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)\left(1 - \sigma(a_n)\right) = y_n(1 - y_n). \tag{161}$$

Finally, we have

$$\nabla a_n = \boldsymbol{\phi}_n \tag{162}$$

where ∇ denotes the gradient with respect to $\mathbf{w}$. Combining (160), (161) and (162) using the chain rule, we obtain

$$\begin{aligned} \nabla E &= \sum_{n=1}^{N} \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial a_n}\nabla a_n \\ &= \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi}_n \end{aligned}$$

as required.

4.14 If the data set is linearly separable, any decision boundary separating the two classes will have the property

$$\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n \begin{cases} \geqslant 0 & \text{if } t_n = 1, \\ < 0 & \text{otherwise.} \end{cases}$$

Moreover, from (4.90) we see that the negative log-likelihood will be minimized (i.e., the likelihood maximized) when $y_n = \sigma\left(\mathbf{w}_{\mathrm{T}}\boldsymbol{\phi}_n\right) = t_n$ for all n. This will be the case when the sigmoid function is saturated, which occurs when its argument, $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$, goes to $\pm\infty$, i.e., when the magnitude of $\mathbf{w}$ goes to infinity.

4.15 **NOTE**: In PRML, "concave" should be "convex" on the last line of the exercise.

Assuming that the argument to the sigmoid function (4.87) is finite, the diagonal elements of $\mathbf{R}$ will be strictly positive. Then

$$\mathbf{v}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\mathbf{v} = \left(\mathbf{v}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}^{1/2}\right)\left(\mathbf{R}^{1/2}\boldsymbol{\Phi}\mathbf{v}\right) = \left\|\mathbf{R}^{1/2}\boldsymbol{\Phi}\mathbf{v}\right\|^2 > 0$$

where $\mathbf{R}^{1/2}$ is a diagonal matrix with elements $(y_n(1 - y_n))^{1/2}$, and thus $\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}$ is positive definite.

Now consider a Taylor expansion of $E(\mathbf{w})$ around a minima, $\mathbf{w}^\star$,

$$E(\mathbf{w}) = E(\mathbf{w}^\star) + \frac{1}{2}\left(\mathbf{w} - \mathbf{w}^\star\right)^{\mathrm{T}}\mathbf{H}\left(\mathbf{w} - \mathbf{w}^\star\right)$$

where the linear term has vanished since $\mathbf{w}^\star$ is a minimum. Now let

$$\mathbf{w} = \mathbf{w}^\star + \lambda\mathbf{v}$$

where $\mathbf{v}$ is an arbitrary, non-zero vector in the weight space and consider

$$\frac{\partial^2 E}{\partial \lambda^2} = \mathbf{v}^{\mathrm{T}} \mathbf{H} \mathbf{v} > 0.$$

This shows that $E(\mathbf{w})$ is convex. Moreover, at the minimum of $E(\mathbf{w})$,

$$\mathbf{H}\left(\mathbf{w} - \mathbf{w}^{\star}\right) = 0$$

and since $\mathbf{H}$ is positive definite, $\mathbf{H}^{-1}$ exists and $\mathbf{w} = \mathbf{w}^{\star}$ must be the unique minimum.

4.16 If the values of the $\{t_n\}$ were known then each data point for which $t_n = 1$ would contribute $p(t_n = 1|\boldsymbol{\phi}(\mathbf{x}_n))$ to the log likelihood, and each point for which $t_n = 0$ would contribute $1 - p(t_n = 1|\boldsymbol{\phi}(\mathbf{x}_n))$ to the log likelihood. A data point whose probability of having $t_n = 1$ is given by π_n will therefore contribute

$$\pi_n p(t_n = 1|\boldsymbol{\phi}(\mathbf{x}_n)) + (1 - \pi_n)(1 - p(t_n = 1|\boldsymbol{\phi}(\mathbf{x}_n)))$$

and so the overall log likelihood for the data set is given by

$$\sum_{n=1}^{N} \pi_n \ln p\left(t_n = 1 \mid \boldsymbol{\phi}(\mathbf{x}_n)\right) + (1 - \pi_n) \ln\left(1 - p\left(t_n = 1 \mid \boldsymbol{\phi}(\mathbf{x}_n)\right)\right). \tag{163}$$

This can also be viewed from a sampling perspective by imagining sampling the value of each t_n some number M times, with probability of $t_n = 1$ given by π_n, and then constructing the likelihood function for this expanded data set, and dividing by M. In the limit $M \to \infty$ we recover (163).

4.17 From (4.104) we have

$$\begin{aligned}
\frac{\partial y_k}{\partial a_k} &= \frac{e^{a_k}}{\sum_i e^{a_i}} - \left(\frac{e^{a_k}}{\sum_i e^{a_i}}\right)^2 = y_k(1 - y_k), \\
\frac{\partial y_k}{\partial a_j} &= -\frac{e^{a_k} e^{a_j}}{\left(\sum_i e^{a_i}\right)^2} = -y_k y_j, \qquad j \neq k.
\end{aligned}$$

Combining these results we obtain (4.106).

4.18 **NOTE**: In the 1st printing of PRML, the text of the exercise refers equation (4.91) where it should refer to (4.106).

From (4.108) we have

$$\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}.$$

If we combine this with (4.106) using the chain rule, we get

$$\begin{aligned}\frac{\partial E}{\partial a_{nj}} &= \sum_{k=1}^{K} \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} \\ &= -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk} \left(I_{kj} - y_{nj}\right) \\ &= y_{nj} - t_{nj},\end{aligned}$$

where we have used that $\forall n : \sum_k t_{nk} = 1$.

If we combine this with (162), again using the chain rule, we obtain (4.109).

4.19 Using the cross-entropy error function (4.90), and following Exercise 4.13, we have

$$\frac{\partial E}{\partial y_n} = \frac{y_n - t_n}{y_n(1 - y_n)}. \tag{164}$$

Also

$$\nabla a_n = \boldsymbol{\phi}_n. \tag{165}$$

From (4.115) and (4.116) we have

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \Phi(a_n)}{\partial a_n} = \frac{1}{\sqrt{2\pi}} e^{-a_n^2}. \tag{166}$$

Combining (164), (165) and (166), we get

$$\nabla E = \sum_{n=1}^{N} \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial a_n} \nabla a_n = \sum_{n=1}^{N} \frac{y_n - t_n}{y_n(1 - y_n)} \frac{1}{\sqrt{2\pi}} e^{-a_n^2} \boldsymbol{\phi}_n. \tag{167}$$

In order to find the expression for the Hessian, it is is convenient to first determine

$$\begin{aligned}\frac{\partial}{\partial y_n} \frac{y_n - t_n}{y_n(1 - y_n)} &= \frac{y_n(1 - y_n)}{y_n^2(1 - y_n)^2} - \frac{(y_n - t_n)(1 - 2y_n)}{y_n^2(1 - y_n)^2} \\ &= \frac{y_n^2 + t_n - 2y_n t_n}{y_n^2(1 - y_n)^2}.\end{aligned} \tag{168}$$

Then using (165)–(168) we have

$$\begin{aligned}\nabla\nabla E &= \sum_{n=1}^{N} \left\{ \frac{\partial}{\partial y_n} \left[\frac{y_n - t_n}{y_n(1 - y_n)} \right] \frac{1}{\sqrt{2\pi}} e^{-a_n^2} \boldsymbol{\phi}_n \nabla y_n \right. \\ &\quad \left. + \frac{y_n - t_n}{y_n(1 - y_n)} \frac{1}{\sqrt{2\pi}} e^{-a_n^2} (-2a_n) \boldsymbol{\phi}_n \nabla a_n \right\} \\ &= \sum_{n=1}^{N} \left(\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1 - y_n)} \frac{1}{\sqrt{2\pi}} e^{-a_n^2} - 2a_n(y_n - t_n) \right) \frac{e^{-2a_n^2} \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}}{\sqrt{2\pi} y_n(1 - y_n)}.\end{aligned}$$

4.20 **NOTE**: In the 1st printing of PRML, equation (4.110) contains an incorrect leading minus sign ('$-$') on the right hand side.

We first write out the components of the $MK \times MK$ Hessian matrix in the form

$$\frac{\partial^2 E}{\partial w_{ki} \partial w_{jl}} = \sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj})\phi_{ni}\phi_{nl}.$$

To keep the notation uncluttered, consider just one term in the summation over n and show that this is positive semi-definite. The sum over n will then also be positive semi-definite. Consider an arbitrary vector of dimension MK with elements u_{ki}. Then

$$\begin{aligned}
\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} &= \sum_{i,j,k,l} u_{ki} y_k (I_{kj} - y_j)\phi_i \phi_l u_{jl} \\
&= \sum_{j,k} b_j y_k (I_{kj} - y_j) b_k \\
&= \sum_{k} y_k b_k^2 - \left(\sum_k b_k y_k \right)^2
\end{aligned}$$

where

$$b_k = \sum_i u_{ki}\phi_{ni}.$$

We now note that the quantities y_k satisfy $0 \leqslant y_k \leqslant 1$ and $\sum_k y_k = 1$. Furthermore, the function $f(b) = b^2$ is a concave function. We can therefore apply Jensen's inequality to give

$$\sum_k y_k b_k^2 = \sum_k y_k f(b_k) \geqslant f\left(\sum_k y_k b_k \right) = \left(\sum_k y_k b_k \right)^2$$

and hence

$$\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} \geqslant 0.$$

Note that the equality will never arise for finite values of a_k where a_k is the set of arguments to the softmax function. However, the Hessian can be positive *semi*-definite since the basis vectors ϕ_{ni} could be such as to have zero dot product for a linear subspace of vectors u_{ki}. In this case the minimum of the error function would comprise a continuum of solutions all having the same value of the error function.

4.21 **NOTE**: In PRML, (4.116) should read

$$\Phi(a) = \frac{1}{2}\left\{ 1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}} \right) \right\}.$$

Note that $\mathbf{\Phi}$ should be Φ (i.e. not bold) on the l.h.s.

We consider the two cases where $a \geqslant 0$ and $a < 0$ separately. In the first case, we can use (2.42) to rewrite (4.114) as

$$\begin{aligned}\Phi(a) &= \int_{-\infty}^{0} \mathcal{N}\left(\theta|0,1\right)\,\mathrm{d}\theta + \int_{0}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right)\,\mathrm{d}\theta \\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{a/\sqrt{2}} \exp\left(-u^2\right)\sqrt{2}\,\mathrm{d}u \\ &= \frac{1}{2}\left\{1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right\},\end{aligned}$$

where, in the last line, we have used (4.115).

When $a < 0$, the symmetry of the Gaussian distribution gives

$$\Phi(a) = 1 - \Phi(-a).$$

Combining this with the above result, we get

$$\begin{aligned}\Phi(a) &= 1 - \frac{1}{2}\left\{1 + \mathrm{erf}\left(-\frac{a}{\sqrt{2}}\right)\right\} \\ &= \frac{1}{2}\left\{1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right\},\end{aligned}$$

where we have used the fact that the erf function is is anti-symmetric, i.e., $\mathrm{erf}(-a) = -\mathrm{erf}(a)$.

4.22 Starting from (4.136), using (4.135), we have

$$\begin{aligned}p\left(\mathcal{D}\right) &= \int p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)\,\mathrm{d}\boldsymbol{\theta} \\ &\simeq p\left(\mathcal{D} \mid \boldsymbol{\theta}_{\mathrm{MAP}}\right) p\left(\boldsymbol{\theta}_{\mathrm{MAP}}\right) \\ &\qquad \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}})\mathbf{A}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}})\right)\,\mathrm{d}\boldsymbol{\theta} \\ &= p\left(\mathcal{D} \mid \boldsymbol{\theta}_{\mathrm{MAP}}\right) p\left(\boldsymbol{\theta}_{\mathrm{MAP}}\right) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}},\end{aligned}$$

where $\mathbf{A}$ is given by (4.138). Taking the logarithm of this yields (4.137).

4.23 **NOTE**: In the 1st printing of PRML, the text of the exercise contains a typographical error. Following the equation, it should say that $\mathbf{H}$ is the matrix of second derivatives of the *negative* log likelihood.

The BIC approximation can be viewed as a large N approximation to the log model evidence. From (4.138), we have

$$\begin{aligned}\mathbf{A} &= -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) \\ &= \mathbf{H} - \nabla\nabla \ln p(\boldsymbol{\theta}_{\mathrm{MAP}})\end{aligned}$$

and if $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0)$, this becomes

$$\mathbf{A} = \mathbf{H} + \mathbf{V}_0^{-1}.$$

If we assume that the prior is broad, or equivalently that the number of data points is large, we can neglect the term $\mathbf{V}_0^{-1}$ compared to $\mathbf{H}$. Using this result, (4.137) can be rewritten in the form

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const} \tag{169}$$

as required. Note that the phrasing of the question is misleading, since the assumption of a broad prior, or of large N, is required in order to derive this form, as well as in the subsequent simplification.

We now again invoke the broad prior assumption, allowing us to neglect the second term on the right hand side of (169) relative to the first term.

Since we assume i.i.d. data, $\mathbf{H} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})$ consists of a sum of terms, one term for each datum, and we can consider the following approximation:

$$\mathbf{H} = \sum_{n=1}^{N} \mathbf{H}_n = N\widehat{\mathbf{H}}$$

where $\mathbf{H}_n$ is the contribution from the n^{th} data point and

$$\widehat{\mathbf{H}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{H}_n.$$

Combining this with the properties of the determinant, we have

$$\ln|\mathbf{H}| = \ln|N\widehat{\mathbf{H}}| = \ln\left(N^M|\widehat{\mathbf{H}}|\right) = M\ln N + \ln|\widehat{\mathbf{H}}|$$

where M is the dimensionality of $\boldsymbol{\theta}$. Note that we are assuming that $\widehat{\mathbf{H}}$ has full rank M. Finally, using this result together (169), we obtain (4.139) by dropping the $\ln|\widehat{\mathbf{H}}|$ since this $O(1)$ compared to $\ln N$.

4.24 Consider a rotation of the coordinate axes of the M-dimensional vector $\mathbf{w}$ such that $\mathbf{w} = (w_\parallel, \mathbf{w}_\perp)$ where $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi} = w_\parallel\|\boldsymbol{\phi}\|$, and $\mathbf{w}_\perp$ is a vector of length $M-1$. We then have

$$\begin{aligned}\int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w} &= \iint \sigma\left(w_\parallel\|\boldsymbol{\phi}\|\right) q(\mathbf{w}_\perp|w_\parallel)q(w_\parallel)\,\mathrm{d}w_\parallel\,\mathrm{d}\mathbf{w}_\perp \\ &= \int \sigma(w_\parallel\|\boldsymbol{\phi}\|)q(w_\parallel)\,\mathrm{d}w_\parallel.\end{aligned}$$

Note that the joint distribution $q(\mathbf{w}_\perp, w_\parallel)$ is Gaussian. Hence the marginal distribution $q(w_\parallel)$ is also Gaussian and can be found using the standard results presented in

Section 2.3.2. Denoting the unit vector

$$\mathbf{e} = \frac{1}{\|\boldsymbol{\phi}\|}\boldsymbol{\phi}$$

we have

$$q(w_\parallel) = \mathcal{N}(w_\parallel|\mathbf{e}^{\mathrm{T}}\mathbf{m}_N, \mathbf{e}^{\mathrm{T}}\mathbf{S}_N\mathbf{e}).$$

Defining $a = w_\parallel\|\boldsymbol{\phi}\|$ we see that the distribution of a is given by a simple re-scaling of the Gaussian, so that

$$q(a) = \mathcal{N}(a|\boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N, \boldsymbol{\phi}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi})$$

where we have used $\|\boldsymbol{\phi}\|\mathbf{e} = \boldsymbol{\phi}$. Thus we obtain (4.151) with μ_a given by (4.149) and σ_a^2 given by (4.150).

4.25 From (4.88) we have that

$$\begin{aligned}\left.\frac{\mathrm{d}\sigma}{\mathrm{d}a}\right|_{a=0} &= \sigma(0)(1-\sigma(0)) \\ &= \frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4}. \end{aligned} \tag{170}$$

Since the derivative of a cumulative distribution function is simply the corresponding density function, (4.114) gives

$$\begin{aligned}\left.\frac{\mathrm{d}\Phi(\lambda a)}{\mathrm{d}a}\right|_{a=0} &= \lambda\mathcal{N}\left(0|0,1\right) \\ &= \lambda\frac{1}{\sqrt{2\pi}}.\end{aligned}$$

Setting this equal to (170), we see that

$$\lambda = \frac{\sqrt{2\pi}}{4} \quad \text{or equivalently} \quad \lambda^2 = \frac{\pi}{8}.$$

This is illustrated in Figure 4.9.

4.26 First of all consider the derivative of the right hand side with respect to μ, making use of the definition of the probit function, giving

$$\left(\frac{1}{2\pi}\right)^{1/2}\exp\left\{-\frac{\mu^2}{2(\lambda^{-2}+\sigma^2)}\right\}\frac{1}{(\lambda^{-2}+\sigma^2)^{1/2}}.$$

Now make the change of variable $a = \mu + \sigma z$, so that the left hand side of (4.152) becomes

$$\int_{-\infty}^{\infty}\Phi(\lambda\mu+\lambda\sigma z)\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2}z^2\right\}\sigma\,\mathrm{d}z$$

where we have substituted for the Gaussian distribution. Now differentiate with respect to μ, making use of the definition of the probit function, giving

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}z^2-\frac{\lambda^2}{2}(\mu+\sigma z)^2\right\}\sigma\,\mathrm{d}z.$$

The integral over z takes the standard Gaussian form and can be evaluated analytically by making use of the standard result for the normalization coefficient of a Gaussian distribution. To do this we first complete the square in the exponent

$$\begin{aligned}
&-\frac{1}{2}z^2-\frac{\lambda^2}{2}(\mu+\sigma z)^2\\
&\quad=\quad -\frac{1}{2}z^2(1+\lambda^2\sigma^2)-z\lambda^2\mu\sigma-\frac{1}{2}\lambda^2\mu^2\\
&\quad=\quad -\frac{1}{2}\left[z+\lambda^2\mu\sigma(1+\lambda^2\sigma^2)^{-1}\right]^2(1+\lambda^2\sigma^2)+\frac{1}{2}\frac{\lambda^4\mu^2\sigma^2}{(1+\lambda^2\sigma^2)}-\frac{1}{2}\lambda^2\mu^2.
\end{aligned}$$

Integrating over z then gives the following result for the derivative of the left hand side

$$\begin{aligned}
&\frac{1}{(2\pi)^{1/2}}\frac{1}{(1+\lambda^2\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2}\lambda^2\mu^2+\frac{1}{2}\frac{\lambda^4\mu^2\sigma^2}{(1+\lambda^2\sigma^2)}\right\}\\
&\quad=\quad \frac{1}{(2\pi)^{1/2}}\frac{1}{(1+\lambda^2\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2}\frac{\lambda^2\mu^2}{(1+\lambda^2\sigma^2)}\right\}.
\end{aligned}$$

Thus the derivatives of the left and right hand sides of (4.152) with respect to μ are equal. It follows that the left and right hand sides are equal up to a function of σ^2 and λ. Taking the limit $\mu\to-\infty$ the left and right hand sides both go to zero, showing that the constant of integration must also be zero.

Chapter 5 Neural Networks

5.1 **NOTE**: In the 1st printing of PRML, the text of this exercise contains a typographical error. On line 2, $g(\cdot)$ should be replaced by $h(\cdot)$.

See Solution 3.1.

5.2 The likelihood function for an i.i.d. data set, $\{(\mathbf{x}_1,\mathbf{t}_1),\ldots,(\mathbf{x}_N,\mathbf{t}_N)\}$, under the conditional distribution (5.16) is given by

$$\prod_{n=1}^{N}\mathcal{N}\left(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n,\mathbf{w}),\beta^{-1}\mathbf{I}\right).$$

If we take the logarithm of this, using (2.43), we get

$$\begin{aligned}
&\sum_{n=1}^{N} \ln \mathcal{N}\left(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}\right) \\
&= -\frac{1}{2}\sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^{\mathrm{T}} (\beta\mathbf{I}) \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right) + \text{const} \\
&= -\frac{\beta}{2}\sum_{n=1}^{N} \|\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\|^2 + \text{const},
\end{aligned}$$

where 'const' comprises terms which are independent of $\mathbf{w}$. The first term on the right hand side is proportional to the negative of (5.11) and hence maximizing the log-likelihood is equivalent to minimizing the sum-of-squares error.

5.3 In this case, the likelihood function becomes

$$p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{w}), \boldsymbol{\Sigma}\right),$$

with the corresponding log-likelihood function

$$\begin{aligned}
&\ln p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Sigma}) \\
&\quad = -\frac{N}{2}\left(\ln|\boldsymbol{\Sigma}| + K\ln(2\pi)\right) - \frac{1}{2}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{y}_n), \quad (171)
\end{aligned}$$

where $\mathbf{y}_n = \mathbf{y}(\mathbf{x}_n, \mathbf{w})$ and K is the dimensionality of $\mathbf{y}$ and $\mathbf{t}$.

If we first treat $\boldsymbol{\Sigma}$ as fixed and known, we can drop terms that are independent of $\mathbf{w}$ from (171), and by changing the sign we get the error function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{y}_n).$$

If we consider maximizing (171) w.r.t. $\boldsymbol{\Sigma}$, the terms that need to be kept are

$$-\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{y}_n).$$

By rewriting the second term we get

$$-\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}_n)(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}\right].$$

Using results from Appendix C, we can maximize this by setting the derivative w.r.t. $\mathbf{\Sigma}^{-1}$ to zero, yielding

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}.$$

Thus the optimal value for $\mathbf{\Sigma}$ depends on $\mathbf{w}$ through $\mathbf{y}_n$.

A possible way to address this mutual dependency between $\mathbf{w}$ and $\mathbf{\Sigma}$ when it comes to optimization, is to adopt an iterative scheme, alternating between updates of $\mathbf{w}$ and $\mathbf{\Sigma}$ until some convergence criterion is reached.

5.4 Let $t \in \{0, 1\}$ denote the data set label and let $k \in \{0, 1\}$ denote the true class label. We want the network output to have the interpretation $y(\mathbf{x}, \mathbf{w}) = p(k = 1|\mathbf{x})$. From the rules of probability we have

$$p(t = 1|\mathbf{x}) = \sum_{k=0}^{1} p(t = 1|k)p(k|\mathbf{x}) = (1 - \epsilon)y(\mathbf{x}, \mathbf{w}) + \epsilon(1 - y(\mathbf{x}, \mathbf{w})).$$

The conditional probability of the data label is then

$$p(t|\mathbf{x}) = p(t = 1|\mathbf{x})^t(1 - p(t = 1|\mathbf{x})^{1-t}.$$

Forming the likelihood and taking the negative logarithm we then obtain the error function in the form

$$\begin{aligned} E(\mathbf{w}) \quad = \quad & -\sum_{n=1}^{N}\{t_n \ln[(1 - \epsilon)y(\mathbf{x}_n, \mathbf{w}) + \epsilon(1 - y(\mathbf{x}_n, \mathbf{w}))] \\ & +(1 - t_n)\ln[1 - (1 - \epsilon)y(\mathbf{x}_n, \mathbf{w}) - \epsilon(1 - y(\mathbf{x}_n, \mathbf{w}))]\}. \end{aligned}$$

See also Solution 4.16.

5.5 For the given interpretation of $y_k(\mathbf{x}, \mathbf{w})$, the conditional distribution of the target vector for a multiclass neural network is

$$p(\mathbf{t}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{k=1}^{K} y_k^{t_k}.$$

Thus, for a data set of N points, the likelihood function will be

$$p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) \quad = \quad \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}.$$

Taking the negative logarithm in order to derive an error function we obtain (5.24) as required. Note that this is the same result as for the multiclass logistic regression model, given by (4.108) .

5.6 Differentiating (5.21) with respect to the activation a_n corresponding to a particular data point n, we obtain

$$\frac{\partial E}{\partial a_n} = -t_n \frac{1}{y_n}\frac{\partial y_n}{\partial a_n} + (1-t_n)\frac{1}{1-y_n}\frac{\partial y_n}{\partial a_n}. \tag{172}$$

From (4.88), we have

$$\frac{\partial y_n}{\partial a_n} = y_n(1-y_n). \tag{173}$$

Substituting (173) into (172), we get

$$\begin{aligned}\frac{\partial E}{\partial a_n} &= -t_n \frac{y_n(1-y_n)}{y_n} + (1-t_n)\frac{y_n(1-y_n)}{(1-y_n)} \\ &= y_n - t_n\end{aligned}$$

as required.

5.7 See Solution 4.17.

5.8 From (5.59), using standard derivatives, we get

$$\begin{aligned}\frac{\mathrm{d}\tanh}{\mathrm{d}a} &= \frac{e^a}{e^a+e^{-a}} - \frac{e^a(e^a-e^{-a})}{(e^a+e^{-a})^2} + \frac{e^{-a}}{e^a+e^{-a}} + \frac{e^{-a}(e^a-e^{-a})}{(e^a+e^{-a})^2} \\ &= \frac{e^a+e^{-a}}{e^a+e^{-a}} + \frac{1-e^{2a}-e^{-2a}+1}{(e^a+e^{-a})^2} \\ &= 1 - \frac{e^{2a}-2+e^{-2a}}{(e^a+e^{-a})^2} \\ &= 1 - \frac{(e^a-e^{-a})(e^a-e^{-a})}{(e^a+e^{-a})\,(e^a+e^{-a})} \\ &= 1-\tanh^2(a)\end{aligned}$$

5.9 This simply corresponds to a scaling and shifting of the binary outputs, which directly gives the activation function, using the notation from (5.19), in the form

$$y = 2\sigma(a) - 1.$$

The corresponding error function can be constructed from (5.21) by applying the inverse transform to y_n and t_n, yielding

$$\begin{aligned}E(\mathbf{w}) &= -\sum_{n=1}^{N} \frac{1+t_n}{2}\ln\frac{1+y_n}{2} + \left(1-\frac{1+t_n}{2}\right)\ln\left(1-\frac{1+y_n}{2}\right) \\ &= -\frac{1}{2}\sum_{n=1}^{N}\{(1+t_n)\ln(1+y_n) + (1-t_n)\ln(1-y_n)\} + N\ln 2\end{aligned}$$

where the last term can be dropped, since it is independent of $\mathbf{w}$.

To find the corresponding activation function we simply apply the linear transformation to the logistic sigmoid given by (5.19), which gives

$$\begin{aligned} y(a) &= 2\sigma(a) - 1 = \frac{2}{1+e^{-a}} - 1 \\ &= \frac{1-e^{-a}}{1+e^{-a}} = \frac{e^{a/2} - e^{-a/2}}{e^{a/2} + e^{-a/2}} \\ &= \tanh(a/2). \end{aligned}$$

5.10 From (5.33) and (5.35) we have

$$\mathbf{u}_i^{\mathrm{T}} \mathbf{H} \mathbf{u}_i = \mathbf{u}_i^{\mathrm{T}} \lambda_i \mathbf{u}_i = \lambda_i.$$

Assume that $\mathbf{H}$ is positive definite, so that (5.37) holds. Then by setting $\mathbf{v} = \mathbf{u}_i$ it follows that

$$\lambda_i = \mathbf{u}_i^{\mathrm{T}} \mathbf{H} \mathbf{u}_i > 0 \tag{174}$$

for all values of i. Thus, if $\mathbf{H}$ is positive definite, all of its eigenvalues will be positive.

Conversely, assume that (174) holds. Then, for any vector, $\mathbf{v}$, we can make use of (5.38) to give

$$\begin{aligned} \mathbf{v}^{\mathrm{T}} \mathbf{H} \mathbf{v} &= \left(\sum_i c_i \mathbf{u}_i \right)^{\mathrm{T}} \mathbf{H} \left(\sum_j c_j \mathbf{u}_j \right) \\ &= \left(\sum_i c_i \mathbf{u}_i \right)^{\mathrm{T}} \left(\sum_j \lambda_j c_j \mathbf{u}_j \right) \\ &= \sum_i \lambda_i c_i^2 > 0 \end{aligned}$$

where we have used (5.33) and (5.34) along with (174). Thus, if all of the eigenvalues are positive, the Hessian matrix will be positive definite.

5.11 **NOTE**: In PRML, Equation (5.32) contains a typographical error: $=$ should be $\simeq$.

We start by making the change of variable given by (5.35) which allows the error function to be written in the form (5.36). Setting the value of the error function $E(\mathbf{w})$ to a constant value C we obtain

$$E(\mathbf{w}^\star) + \frac{1}{2} \sum_i \lambda_i \alpha_i^2 = C.$$

Re-arranging gives

$$\sum_i \lambda_i \alpha_i^2 = 2C - 2E(\mathbf{w}^\star) = \widetilde{C}$$

where $\widetilde{C}$ is also a constant. This is the equation for an ellipse whose axes are aligned with the coordinates described by the variables $\{\alpha_i\}$. The length of axis j is found by setting $\alpha_i = 0$ for all $i \neq j$, and solving for α_j giving

$$\alpha_j = \left(\frac{\widetilde{C}}{\lambda_j}\right)^{1/2}$$

which is inversely proportional to the square root of the corresponding eigenvalue.

5.12 **NOTE**: See note in Solution 5.11.

From (5.37) we see that, if $\mathbf{H}$ is positive definite, then the second term in (5.32) will be positive whenever $(\mathbf{w} - \mathbf{w}^\star)$ is non-zero. Thus the smallest value which $E(\mathbf{w})$ can take is $E(\mathbf{w}^\star)$, and so $\mathbf{w}^\star$ is the minimum of $E(\mathbf{w})$.

Conversely, if $\mathbf{w}^\star$ is the minimum of $E(\mathbf{w})$, then, for any vector $\mathbf{w} \neq \mathbf{w}^\star$, $E(\mathbf{w}) > E(\mathbf{w}^\star)$. This will only be the case if the second term of (5.32) is positive for all values of $\mathbf{w} \neq \mathbf{w}^\star$ (since the first term is independent of $\mathbf{w}$). Since $\mathbf{w} - \mathbf{w}^\star$ can be set to any vector of real numbers, it follows from the definition (5.37) that $\mathbf{H}$ must be positive definite.

5.13 From exercise 2.21 we know that a $W \times W$ matrix has $W(W+1)/2$ independent elements. Add to that the W elements of the gradient vector $\mathbf{b}$ and we get

$$\frac{W(W+1)}{2} + W = \frac{W(W+1) + 2W}{2} = \frac{W^2 + 3W}{2} = \frac{W(W+3)}{2}.$$

5.14 We are interested in determining how the correction term

$$\delta = E'(w_{ij}) - \frac{E(w_{ij} + \epsilon) - E(w_{ij} - \epsilon)}{2\epsilon} \tag{175}$$

depend on ϵ.

Using Taylor expansions, we can rewrite the numerator of the first term of (175) as

$$\begin{aligned} &E(w_{ij}) + \epsilon E'(w_{ij}) + \frac{\epsilon^2}{2} E''(w_{ij}) + O(\epsilon^3) \\ &\quad - E(w_{ij}) + \epsilon E'(w_{ij}) - \frac{\epsilon^2}{2} E''(w_{ij}) + O(\epsilon^3) = 2\epsilon E'(w_{ij}) + O(\epsilon^3). \end{aligned}$$

Note that the ϵ^2-terms cancel. Substituting this into (175) we get,

$$\delta = \frac{2\epsilon E'(w_{ij}) + O(\epsilon^3)}{2\epsilon} - E'(w_{ij}) = O(\epsilon^2).$$

5.15 The alternative forward propagation scheme takes the first line of (5.73) as its starting point. However, rather than proceeding with a 'recursive' definition of $\partial y_k / \partial a_j$, we instead make use of a corresponding definition for $\partial a_j / \partial x_i$. More formally

$$J_{ki} = \frac{\partial y_k}{\partial x_i} = \sum_j \frac{\partial y_k}{\partial a_j} \frac{\partial a_j}{\partial x_i}$$

where $\partial y_k/\partial a_j$ is defined by (5.75), (5.76) or simply as δ_{kj}, for the case of linear output units. We define $\partial a_j/\partial x_i = w_{ji}$ if a_j is in the first hidden layer and otherwise

$$\frac{\partial a_j}{\partial x_i} = \sum_l \frac{\partial a_j}{\partial a_l}\frac{\partial a_l}{\partial x_i} \tag{176}$$

where

$$\frac{\partial a_j}{\partial a_l} = w_{jl}h'(a_l). \tag{177}$$

Thus we can evaluate J_{ki} by forward propagating $\partial a_j/\partial x_i$, with initial value w_{ij}, alongside a_j, using (176) and (177).

5.16 The multivariate form of (5.82) is

$$E = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}(\mathbf{y}_n - \mathbf{t}_n).$$

The elements of the first and second derivatives then become

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}\frac{\partial \mathbf{y}_n}{\partial w_i}$$

and

$$\frac{\partial^2 E}{\partial w_i \partial w_j} = \sum_{n=1}^{N}\left\{\frac{\partial \mathbf{y}_n}{\partial w_j}^{\mathrm{T}}\frac{\partial \mathbf{y}_n}{\partial w_i} + (\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}\frac{\partial^2 \mathbf{y}_n}{\partial w_j\,\partial w_i}\right\}.$$

As for the univariate case, we again assume that the second term of the second derivative vanishes and we are left with

$$\mathbf{H} = \sum_{n=1}^{N}\mathbf{B}_n\mathbf{B}_n^{\mathrm{T}},$$

where $\mathbf{B}_n$ is a $W \times K$ matrix, K being the dimensionality of $\mathbf{y}_n$, with elements

$$(\mathbf{B}_n)_{lk} = \frac{\partial y_{nk}}{\partial w_l}.$$

5.17 Taking the second derivatives of (5.193) with respect to two weights w_r and w_s we obtain

$$\begin{aligned}\frac{\partial^2 E}{\partial w_r \partial w_s} &= \sum_k \int \left\{\frac{\partial y_k}{\partial w_r}\frac{\partial y_k}{\partial w_s}\right\} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\quad + \sum_k \int \left\{\frac{\partial^2 y_k}{\partial w_r \partial w_s}(y_k(\mathbf{x}) - \mathbb{E}_{t_k}[t_k|\mathbf{x}])\right\} p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \end{aligned} \tag{178}$$

Using the result (1.89) that the outputs $y_k(\mathbf{x})$ of the trained network represent the conditional averages of the target data, we see that the second term in (178) vanishes. The Hessian is therefore given by an integral of terms involving only the products of first derivatives. For a finite data set, we can write this result in the form

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \frac{1}{N} \sum_{n=1}^{N} \sum_k \frac{\partial y_k^n}{\partial w_r} \frac{\partial y_k^n}{\partial w_s}$$

which is identical with (5.84) up to a scaling factor.

5.18 If we introduce skip layer weights, $\mathbf{U}$, into the model described in Section 5.3.2, this will only affect the last of the forward propagation equations, (5.64), which becomes

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j + \sum_{i=1}^{D} u_{ki} x_i.$$

Note that there is no need to include the input bias. The derivative w.r.t. u_{ki} can be expressed using the output $\{\delta_k\}$ of (5.65),

$$\frac{\partial E}{\partial u_{ki}} = \delta_k x_i.$$

5.19 If we take the gradient of (5.21) with respect to $\mathbf{w}$, we obtain

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \frac{\partial E}{\partial a_n} \nabla a_n = \sum_{n=1}^{N} (y_n - t_n) \nabla a_n,$$

where we have used the result proved earlier in the solution to Exercise 5.6. Taking the second derivatives we have

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \left\{ \frac{\partial y_n}{\partial a_n} \nabla a_n \nabla a_n + (y_n - t_n) \nabla\nabla a_n \right\}.$$

Dropping the last term and using the result (4.88) for the derivative of the logistic sigmoid function, proved in the solution to Exercise 4.12, we finally get

$$\nabla\nabla E(\mathbf{w}) \simeq \sum_{n=1}^{N} y_n(1 - y_n) \nabla a_n \nabla a_n = \sum_{n=1}^{N} y_n(1 - y_n) \mathbf{b}_n \mathbf{b}_n^{\mathrm{T}}$$

where $\mathbf{b}_n \equiv \nabla a_n$.

5.20 Using the chain rule, we can write the first derivative of (5.24) as

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\partial E}{\partial a_{nk}} \frac{\partial a_{nk}}{\partial w_i}. \tag{179}$$

From Exercise 5.7, we know that

$$\frac{\partial E}{\partial a_{nk}} = y_{nk} - t_{nk}.$$

Using this and (4.106), we can get the derivative of (179) w.r.t. w_j as

$$\frac{\partial^2 E}{\partial w_i \, \partial w_j} = \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\sum_{l=1}^{K} y_{nk}(I_{kl} - y_{nl}) \frac{\partial a_{nk}}{\partial w_i} \frac{\partial a_{nl}}{\partial w_j} + (y_{nk} - t_{nk}) \frac{\partial^2 a_{nk}}{\partial w_i \, \partial w_j} \right).$$

For a trained model, the network outputs will approximate the conditional class probabilities and so the last term inside the parenthesis will vanish in the limit of a large data set, leaving us with

$$(\mathbf{H})_{ij} \simeq \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{K} y_{nk}(I_{kl} - y_{nl}) \frac{\partial a_{nk}}{\partial w_i} \frac{\partial a_{nl}}{\partial w_j}.$$

5.21 **NOTE**: In PRML, the text in the exercise could be misunderstood; a clearer formulation is: "Extend the expression (5.86) for the outer product approximation of the Hessian matrix to the case of $K > 1$ output units. Hence, derive a form that allows (5.87) to be used to incorporate sequentially contributions from individual outputs as well as individual patterns. This, together with the identity (5.88), will allow the use of (5.89) for finding the inverse of the Hessian by sequentially incorporating contributions from individual outputs and patterns."

From (5.44) and (5.46), we see that the multivariate form of (5.82) is

$$E = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} (y_{nk} - t_{nk})^2 .$$

Consequently, the multivariate form of (5.86) is given by

$$\mathbf{H}_{NK} = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbf{b}_{nk} \mathbf{b}_{nk}^{\mathrm{T}} \tag{180}$$

where $\mathbf{b}_{nk} \equiv \nabla a_{nk} = \nabla y_{nk}$. The double index indicate that we will now iterate over outputs as well as patterns in the sequential build-up of the Hessian. However, in terms of the end result, there is no real need to attribute terms in this sum to specific outputs or specific patterns. Thus, by changing the indexation in (180), we can write it

$$\mathbf{H}_J = \sum_{j=1}^{J} \mathbf{c}_j \mathbf{c}_j^{\mathrm{T}} \tag{181}$$

where $J = NK$ and

$$\begin{aligned} \mathbf{c}_j &= \mathbf{b}_{n(j)k(j)} \\ n(j) &= (j-1) \oslash K + 1 \\ k(j) &= (j-1) \odot K + 1 \end{aligned}$$

with $\oslash$ and $\odot$ denoting integer division and remainder, respectively. The advantage of the indextion in (181) is that now we have a single indexed sum and so we can use (5.87)–(5.89) as they stand, just replacing $\mathbf{b}_L$ with $\mathbf{c}_L$, letting L run from 0 to $J-1$.

5.22 **NOTE**: The first printing of PRML contained typographical errors in equation (5.95). On the r.h.s., $H_{kk'}$ should be $M_{kk'}$. Moreover, the indices j and j' should be swapped on the r.h.s.

Using the chain rule together with (5.48) and (5.92), we have

$$\begin{aligned}\frac{\partial E_n}{\partial w_{kj}^{(2)}} &= \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}^{(2)}} \\ &= \delta_k z_j \end{aligned} \tag{182}$$

Thus,

$$\frac{\partial^2 E_n}{\partial w_{kj}^{(2)}\,\partial w_{k'j'}^{(2)}} = \frac{\partial \delta_k z_j}{\partial w_{k'j'}^{(2)}}$$

and since z_j is independent of the second layer weights,

$$\begin{aligned}\frac{\partial^2 E_n}{\partial w_{kj}^{(2)}\,\partial w_{k'j'}^{(2)}} &= z_j \frac{\partial \delta_k}{\partial w_{k'j'}^{(2)}} \\ &= z_j \frac{\partial^2 E_n}{\partial a_k\,\partial a_{k'}}\frac{\partial a_k}{\partial w_{k'j'}^{(2)}} \\ &= z_j z_{j'} M_{kk'},\end{aligned}$$

where we again have used the chain rule together with (5.48) and (5.92).

If both weights are in the first layer, we again used the chain rule, this time together with (5.48), (5.55) and (5.56), to get

$$\begin{aligned}\frac{\partial E_n}{\partial w_{ji}^{(1)}} &= \frac{\partial E_n}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}^{(1)}} \\ &= x_i \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j} \\ &= x_i\, h'(a_j) \sum_k w_{kj}^{(2)} \delta_k.\end{aligned}$$

Thus we have

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)}\,\partial w_{j'i'}^{(1)}} = \frac{\partial}{\partial w_{j'i'}}\left(x_i\, h'(a_j) \sum_k w_{kj}^{(2)} \delta_k \right).$$

Now we note that x_i and $w_{kj}^{(2)}$ do not depend on $w_{j'i'}^{(1)}$, while $h'(a_j)$ is only affected in the case where $j = j'$. Using these observations together with (5.48), we get

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)}\,\partial w_{j'i'}^{(1)}} = x_i x_{i'} h''(a_j) I_{jj'} \sum_k w_{kj}^{(2)} \delta_k + x_i h'(a_j) \sum_k w_{kj}^{(2)} \frac{\partial \delta_k}{\partial w_{j'i'}^{(1)}}. \tag{183}$$

From (5.48), (5.55), (5.56), (5.92) and the chain rule, we have

$$\begin{aligned}\frac{\partial \delta_k}{\partial w_{j'i'}^{(1)}} &= \sum_{k'} \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \frac{\partial a_{k'}}{\partial a_{j'}} \frac{\partial a_{j'}}{\partial w_{j'i'}^{(1)}} \\ &= x_{i'} h'(a_j) \sum_{k'} w_{k'j'}^{(2)} M_{kk'}. \end{aligned} \tag{184}$$

Substituting this back into (183), we obtain (5.94).

Finally, from (182) we have

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj'}^{(2)}} = \frac{\partial \delta_k z_{j'}}{\partial w_{ji}^{(1)}}.$$

Using (184), we get

$$\begin{aligned}\frac{\partial^2 E_n}{\partial w_{ij}^{(1)} \partial w_{kj'}^{(2)}} &= z_{j'} x_i h'(a_j) \sum_{k'} w_{k'j}^{(2)} M_{kk'} + \delta_k I_{jj'} h'(a_j) x_i \\ &= x_i h'(a_j) \left(\delta_k I_{jj'} + \sum_{k'} w_{k'j}^{(2)} M_{kk'} \right). \end{aligned}$$

5.23 If we introduce skip layer weights into the model discussed in Section 5.4.5, three new cases are added to three already covered in Exercise 5.22.

The first derivative w.r.t. skip layer weight u_{ki} can be written

$$\frac{\partial E_n}{\partial u_{ki}} = \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial u_{ki}} = \frac{\partial E_n}{\partial a_k} x_i. \tag{185}$$

Using this, we can consider the first new case, where both weights are in the skip layer,

$$\begin{aligned}\frac{\partial^2 E_n}{\partial u_{ki} \partial u_{k'i'}} &= \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \frac{\partial a_{k'}}{\partial u_{k'i'}} x_i \\ &= M_{kk'} x_i x_{i'}, \end{aligned}$$

where we have also used (5.92).

When one weight is in the skip layer and the other weight is in the hidden-to-output layer, we can use (185), (5.48) and (5.92) to get

$$\begin{aligned}\frac{\partial^2 E_n}{\partial u_{ki} \partial w_{k'j}^{(2)}} &= \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{k'j}^{(2)}} x_i \\ &= M_{kk'} z_j x_i. \end{aligned}$$

Finally, if one weight is a skip layer weight and the other is in the input-to-hidden layer, (185), (5.48), (5.55), (5.56) and (5.92) together give

$$\begin{aligned}\frac{\partial^2 E_n}{\partial u_{ki}\,\partial w_{ji'}^{(1)}} &= \frac{\partial}{\partial w_{ji'}^{(1)}}\left(\frac{\partial E_n}{\partial a_k}x_i\right)\\ &= \sum_{k'}\frac{\partial^2 E_n}{\partial a_k\,\partial a_{k'}}\frac{\partial a_{k'}}{\partial w_{ji'}^{(1)}}x_i\\ &= x_i x_{i'} h'(a_j)\sum_{k'} M_{kk'} w_{k'j}^{(2)}.\end{aligned}$$

5.24 With the transformed inputs, weights and biases, (5.113) becomes

$$z_j = h\left(\sum_i \widetilde{w}_{ji}\widetilde{x}_i + \widetilde{w}_{j0}\right).$$

Using (5.115)–(5.117), we can rewrite the argument of $h(\cdot)$ on the r.h.s. as

$$\begin{aligned}&\sum_i \frac{1}{a}w_{ji}(ax_i+b) + w_{j0} - \frac{b}{a}\sum_i w_{ji}\\ &= \sum_i w_{ji}x_i + \frac{b}{a}\sum_i w_{ji} + w_{j0} - \frac{b}{a}\sum_i w_{ji}\\ &= \sum_i w_{ji}x_i + w_{j0}.\end{aligned}$$

Similarly, with the transformed outputs, weights and biases, (5.114) becomes

$$\widetilde{y}_k = \sum_i \widetilde{w}_{kj}z_j + \widetilde{w}_{k0}.$$

Using (5.118)–(5.120), we can rewrite this as

$$\begin{aligned}cy_k + d &= \sum_k cw_{kj}z_j + cw_{k0} + d\\ &= c\left(\sum_i w_{kj}z_j + w_{k0}\right) + d.\end{aligned}$$

By subtracting d and subsequently dividing by c on both sides, we recover (5.114) in its original form.

5.25 The gradient of (5.195) is given

$$\nabla E = \mathbf{H}(\mathbf{w} - \mathbf{w}^\star)$$

and hence update formula (5.196) becomes

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^{\star}).$$

Pre-multiplying both sides with $\mathbf{u}_j^{\mathrm{T}}$ we get

$$\begin{aligned} w_j^{(\tau)} &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau)} && (186)\\ &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau-1)} - \rho\mathbf{u}_j^{\mathrm{T}}\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^{\star}) \\ &= w_j^{(\tau-1)} - \rho\eta_j\mathbf{u}_j^{\mathrm{T}}(\mathbf{w} - \mathbf{w}^{\star}) \\ &= w_j^{(\tau-1)} - \rho\eta_j(w_j^{(\tau-1)} - w_j^{\star}), && (187) \end{aligned}$$

where we have used (5.198). To show that

$$w_j^{(\tau)} = \{1 - (1 - \rho\eta_j)^{\tau}\}\, w_j^{\star}$$

for $\tau = 1, 2, \ldots$, we can use proof by induction. For $\tau = 1$, we recall that $\mathbf{w}^{(0)} = \mathbf{0}$ and insert this into (187), giving

$$\begin{aligned} w_j^{(1)} &= w_j^{(0)} - \rho\eta_j(w_j^{(0)} - w_j^{\star}) \\ &= \rho\eta_j w_j^{\star} \\ &= \{1 - (1 - \rho\eta_j)\}\, w_j^{\star}. \end{aligned}$$

Now we assume that the result holds for $\tau = N - 1$ and then make use of (187)

$$\begin{aligned} w_j^{(N)} &= w_j^{(N-1)} - \rho\eta_j(w_j^{(N-1)} - w_j^{\star}) \\ &= w_j^{(N-1)}(1 - \rho\eta_j) + \rho\eta_j w_j^{\star} \\ &= \left\{1 - (1 - \rho\eta_j)^{N-1}\right\} w_j^{\star}(1 - \rho\eta_j) + \rho\eta_j w_j^{\star} \\ &= \left\{(1 - \rho\eta_j) - (1 - \rho\eta_j)^{N}\right\} w_j^{\star} + \rho\eta_j w_j^{\star} \\ &= \left\{1 - (1 - \rho\eta_j)^{N}\right\} w_j^{\star} \end{aligned}$$

as required.

Provided that $|1 - \rho\eta_j| < 1$ then we have $(1 - \rho\eta_j)^{\tau} \to 0$ as $\tau \to \infty$, and hence $\left\{1 - (1 - \rho\eta_j)^{N}\right\} \to 1$ and $\mathbf{w}^{(\tau)} \to \mathbf{w}^{\star}$.

If τ is finite but $\eta_j \gg (\rho\tau)^{-1}$, τ must still be large, since $\eta_j\rho\tau \gg 1$, even though $|1 - \rho\eta_j| < 1$. If τ is large, it follows from the argument above that $w_j^{(\tau)} \simeq w_j^{\star}$.

If, on the other hand, $\eta_j \ll (\rho\tau)^{-1}$, this means that $\rho\eta_j$ must be small, since $\rho\eta_j\tau \ll 1$ and τ is an integer greater than or equal to one. If we expand,

$$(1 - \rho\eta_j)^{\tau} = 1 - \tau\rho\eta_j + O(\rho\eta_j^2)$$

and insert this into (5.197), we get

$$\begin{aligned} |w_j^{(\tau)}| &= |\left\{1-(1-\rho\eta_j)^\tau\right\} w_j^\star| \\ &= |\left\{1-(1-\tau\rho\eta_j+O(\rho\eta_j^2))\right\} w_j^\star| \\ &\simeq \tau\rho\eta_j|w_j^\star| \ll |w_j^\star| \end{aligned}$$

Recall that in Section 3.5.3 we showed that when the regularization parameter (called α in that section) is much larger than one of the eigenvalues (called λ_j in that section) then the corresponding parameter value w_i will be close to zero. Conversely, when α is much smaller than λ_i then w_i will be close to its maximum likelihood value. Thus α is playing an analogous role to $\rho\tau$.

5.26 **NOTE**: In PRML, equation (5.201) should read

$$\Omega_n = \frac{1}{2}\sum_k (\mathcal{G}y_k)^2\Big|_{\mathbf{x}_n}.$$

In this solution, we will indicate dependency on $\mathbf{x}_n$ with a subscript n on relevant symbols.

Substituting the r.h.s. of (5.202) into (5.201) and then using (5.70), we get

$$\Omega_n = \frac{1}{2}\sum_k \left(\sum_i \tau_{ni}\frac{\partial y_{nk}}{\partial x_{ni}}\right)^2 \tag{188}$$

$$= \frac{1}{2}\sum_k \left(\sum_i \tau_{ni} J_{nki}\right)^2 \tag{189}$$

where J_{nki} denoted J_{ki} evaluated at $\mathbf{x}_n$. Summing (189) over n, we get (5.128).

By applying $\mathcal{G}$ from (5.202) to the equations in (5.203) and making use of (5.205) we obtain (5.204). From this, we see that β_{nl} can be written in terms of α_{ni}, which in turn can be written as functions of β_{ni} from the previous layer. For the input layer, using (5.204) and (5.205), we get

$$\begin{aligned} \beta_{nj} &= \sum_i w_{ji}\alpha_{ni} \\ &= \sum_i w_{ji}\mathcal{G}x_{ni} \\ &= \sum_i w_{ji}\sum_{i'}\tau_{ni'}\frac{\partial x_{ni}}{\partial x_{ni'}} \\ &= \sum_i w_{ji}\tau_{ni}. \end{aligned} \tag{190}$$

Thus we see that, starting from (190), $\boldsymbol{\tau}_n$ is propagated forward by subsequent application of the equations in (5.204), yielding the β_{nl} for the output layer, from which Ω_n can be computed using (5.201),

$$\Omega_n = \frac{1}{2}\sum_k \left(\mathcal{G}y_{nk}\right)^2 = \frac{1}{2}\sum_k \alpha_{nk}^2.$$

Considering $\partial\Omega_n/\partial w_{rs}$, we start from (5.201) and make use of the chain rule, together with (5.52), (5.205) and (5.207), to obtain

$$\begin{aligned}\frac{\partial\Omega_n}{\partial w_{rs}} &= \sum_k \left(\mathcal{G}y_{nk}\right)\mathcal{G}\left(\delta_{nkr}z_{ns}\right) \\ &= \sum_k \alpha_{nk}\left(\phi_{nkr}z_{ns} + \delta_{nkr}\alpha_{ns}\right).\end{aligned}$$

The backpropagation formula for computing δ_{nkr} follows from (5.74), which is used in computing the Jacobian matrix, and is given by

$$\delta_{nkr} = h'(a_{nr})\sum_l w_{lr}\delta_{nkl}.$$

Using this together with (5.205) and (5.207), we can obtain backpropagation equations for ϕ_{nkr},

$$\begin{aligned}\phi_{nkr} &= \mathcal{G}\delta_{nkr} \\ &= \mathcal{G}\left(h'(a_{nr})\sum_l w_{lr}\delta_{nkl}\right) \\ &= h''(a_{nr})\beta_{nr}\sum_l w_{lr}\delta_{nkl} + h'(a_{nr})\sum_l w_{lr}\phi_{nkl}.\end{aligned}$$

5.27 If $\mathbf{s}(\mathbf{x},\boldsymbol{\xi}) = \mathbf{x} + \boldsymbol{\xi}$, then

$$\frac{\partial s_k}{\partial \xi_i} = I_{ki}, \text{ i.e., } \frac{\partial \mathbf{s}}{\partial \boldsymbol{\xi}} = \mathbf{I},$$

and since the first order derivative is constant, there are no higher order derivatives. We now make use of this result to obtain the derivatives of y w.r.t. ξ_i:

$$\frac{\partial y}{\partial \xi_i} = \sum_k \frac{\partial y}{\partial s_k}\frac{\partial s_k}{\partial \xi_i} = \frac{\partial y}{\partial s_i} = b_i$$

$$\frac{\partial y}{\partial \xi_i \partial \xi_j} = \frac{\partial b_i}{\partial \xi_j} = \sum_k \frac{\partial b_i}{\partial s_k}\frac{\partial s_k}{\partial \xi_j} = \frac{\partial b_i}{\partial s_j} = B_{ij}$$

Using these results, we can write the expansion of $\widetilde{E}$ as follows:

$$\begin{aligned}\widetilde{E} &= \frac{1}{2}\iiint \{y(\mathbf{x})-t\}^2 p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &+ \iiint \{y(\mathbf{x})-t\}\mathbf{b}^{\mathrm{T}}\boldsymbol{\xi} p(\boldsymbol{\xi})p(t|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &+ \frac{1}{2}\iiint \boldsymbol{\xi}^{\mathrm{T}}\left(\{y(\mathbf{x})-t\}\mathbf{B}+\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\boldsymbol{\xi} p(\boldsymbol{\xi})p(t|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.\end{aligned}$$

The middle term will again disappear, since $\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$ and thus we can write $\widetilde{E}$ on the form of (5.131) with

$$\Omega = \frac{1}{2}\iiint \boldsymbol{\xi}^{\mathrm{T}}\left(\{y(\mathbf{x})-t\}\mathbf{B}+\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\boldsymbol{\xi} p(\boldsymbol{\xi})p(t|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

Again the first term within the parenthesis vanishes to leading order in $\boldsymbol{\xi}$ and we are left with

$$\begin{aligned}\Omega &\simeq \frac{1}{2}\iint \boldsymbol{\xi}^{\mathrm{T}}\left(\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\boldsymbol{\xi} p(\boldsymbol{\xi})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\iint \mathrm{Trace}\left[\left(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}\right)\left(\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\right] p(\boldsymbol{\xi})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\int \mathrm{Trace}\left[\mathbf{I}\left(\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\int \mathbf{b}^{\mathrm{T}}\mathbf{b}p(\mathbf{x})\,\mathrm{d}\mathbf{x} = \frac{1}{2}\int \|\nabla y(\mathbf{x})\|^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x},\end{aligned}$$

where we used the fact that $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}] = \mathbf{I}$.

5.28 The modifications only affect derivatives with respect to weights in the convolutional layer. The units within a feature map (indexed m) have different inputs, but all share a common weight vector, $\mathbf{w}^{(m)}$. Thus, errors $\delta^{(m)}$ from all units within a feature map will contribute to the derivatives of the corresponding weight vector. In this situation, (5.50) becomes

$$\frac{\partial E_n}{\partial w_i^{(m)}} = \sum_j \frac{\partial E_n}{\partial a_j^{(m)}}\frac{\partial a_j^{(m)}}{\partial w_i^{(m)}} = \sum_j \delta_j^{(m)} z_{ji}^{(m)}.$$

Here $a_j^{(m)}$ denotes the activation of the j^{th} unit in the m^{th} feature map, whereas $w_i^{(m)}$ denotes the i^{th} element of the corresponding feature vector and, finally, $z_{ji}^{(m)}$ denotes the i^{th} input for the j^{th} unit in the m^{th} feature map; the latter may be an actual input or the output of a preceding layer.

Note that $\delta_j^{(m)} = \partial E_n/\partial a_j^{(m)}$ will typically be computed recursively from the δs of the units in the following layer, using (5.55). If there are layer(s) preceding the

convolutional layer, the standard backward propagation equations will apply; the weights in the convolutional layer can be treated as if they were independent parameters, for the purpose of computing the δs for the preceding layer's units.

5.29 This is easily verified by taking the derivative of (5.138), using (1.46) and standard derivatives, yielding

$$\frac{\partial \Omega}{\partial w_i} = \frac{1}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)} \sum_j \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \frac{(w_i - \mu_j)}{\sigma^2}.$$

Combining this with (5.139) and (5.140), we immediately obtain the second term of (5.141).

5.30 Since the μ_js only appear in the regularization term, $\Omega(\mathbf{w})$, from (5.139) we have

$$\frac{\partial \widetilde{E}}{\partial \mu_j} = \lambda \frac{\partial \Omega}{\partial \mu_j}. \tag{191}$$

Using (2.42), (5.138) and (5.140) and standard rules for differentiation, we can calculate the derivative of $\Omega(\mathbf{w})$ as follows:

$$\begin{aligned} \frac{\partial \Omega}{\partial \mu_j} &= -\sum_i \frac{1}{\sum_{j'} \pi_{j'} \mathcal{N}\left(w_i|\mu_{j'}, \sigma_{j'}^2\right)} \pi_j \mathcal{N}\left(w_i|\mu_j, \sigma_j^2\right) \frac{w_i - \mu_j}{\sigma_j^2} \\ &= -\sum_i \gamma_j(\mathbf{w}_i) \frac{w_i - \mu_j}{\sigma_j^2}. \end{aligned}$$

Combining this with (191), we get (5.142).

5.31 Following the same line of argument as in Solution 5.30, we need the derivative of $\Omega(\mathbf{w})$ w.r.t. σ_j. Again using (2.42), (5.138) and (5.140) and standard rules for differentiation, we find this to be

$$\begin{aligned} \frac{\partial \Omega}{\partial \sigma_j} &= -\sum_i \frac{1}{\sum_{j'} \pi_{j'} \mathcal{N}\left(w_i|\mu_{j'}, \sigma_{j'}^2\right)} \pi_j \frac{1}{(2\pi)^{1/2}} \left\{ -\frac{1}{\sigma_j^2} \exp\left(-\frac{(w_i - \mu_j)^2}{2\sigma_j^2}\right) \right. \\ &\qquad \left. + \frac{1}{\sigma_j} \exp\left(-\frac{(w_i - \mu_j)^2}{2\sigma_j^2}\right) \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right\} \\ &= \sum_i \gamma_j(w_i) \left\{ \frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right\}. \end{aligned}$$

Combining this with (191), we get (5.143).

5.32 **NOTE**: In the first printing of PRML, there is a leading λ missing on the r.h.s. of equation (5.147). Moreover, in the text of the exercise (last line), the equation of the constraint to be used should read "$\sum_k \gamma_k(w_i) = 1$ for all i".

Equation (5.208) follows from (5.146) in exactly the same way that (4.106) follows from (4.104) in Solution 4.17.

Just as in Solutions 5.30 and 5.31, η_j only affect $\widetilde{E}$ through $\Omega(\mathbf{w})$. However, η_j will affect π_k for all values of k (not just $j = k$). Thus we have

$$\frac{\partial \Omega}{\partial \eta_j} = \sum_k \frac{\partial \Omega}{\partial \pi_k} \frac{\partial \pi_k}{\partial \eta_j}. \tag{192}$$

From (5.138) and (5.140), we get

$$\frac{\partial \Omega}{\partial \pi_k} = -\sum_i \frac{\gamma_k(w_i)}{\pi_k}.$$

Substituting this and (5.208) into (192) yields

$$\begin{aligned} \frac{\partial \Omega}{\partial \eta_j} &= \frac{\partial \widetilde{E}}{\partial \eta_j} = -\sum_k \sum_i \frac{\gamma_k(w_i)}{\pi_k} \{\delta_{jk}\pi_j - \pi_j \pi_k\} \\ &= \sum_i \{\pi_j - \gamma_j(w_i)\}, \end{aligned}$$

where we have used the fact that $\sum_k \gamma_k(w_i) = 1$ for all i.

5.33 From standard trigometric rules we get the position of the end of the first arm,

$$\left(x_1^{(1)}, x_2^{(1)}\right) = (L_1 \cos(\theta_1), L_1 \sin(\theta_1)).$$

Similarly, the position of the end of the second arm relative to the end of the first arm is given by the corresponding equation, with an angle offset of π (see Figure 5.18), which equals a change of sign

$$\begin{aligned} \left(x_1^{(2)}, x_2^{(2)}\right) &= (L_2 \cos(\theta_1 + \theta_2 - \pi), L_1 \sin(\theta_1 + \theta_2 - \pi)) \\ &= -(L_2 \cos(\theta_1 + \theta_2), L_2 \sin(\theta_1 + \theta_2)). \end{aligned}$$

Putting this together, we must also taken into account that θ_2 is measured relative to the first arm and so we get the position of the end of the second arm relative to the attachment point of the first arm as

$$(x_1, x_2) = (L_1 \cos(\theta_1) - L_2 \cos(\theta_1 + \theta_2), L_1 \sin(\theta_1) - L_2 \sin(\theta_1 + \theta_2)).$$

5.34 **NOTE**: In the 1^{st} printing of PRML, the l.h.s. of (5.154) should be replaced with $\gamma_{nk} = \gamma_k(\mathbf{t}_n|\mathbf{x}_n)$. Accordingly, in (5.155) and (5.156), γ_k should be replaced by γ_{nk} and in (5.156), t_l should be t_{nl}.

We start by using the chain rule to write

$$\frac{\partial E_n}{\partial a_k^\pi} = \sum_{j=1}^{K} \frac{\partial E_n}{\partial \pi_j} \frac{\partial \pi_j}{\partial a_k^\pi}. \tag{193}$$

Note that because of the coupling between outputs caused by the softmax activation function, the dependence on the activation of a single output unit involves all the output units.

For the first factor inside the sum on the r.h.s. of (193), standard derivatives applied to the n^{th} term of (5.153) gives

$$\frac{\partial E_n}{\partial \pi_j} = -\frac{\mathcal{N}_{nj}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}} = -\frac{\gamma_{nj}}{\pi_j}. \tag{194}$$

For the for the second factor, we have from (4.106) that

$$\frac{\partial \pi_j}{\partial a_k^\pi} = \pi_j (I_{jk} - \pi_k). \tag{195}$$

Combining (193), (194) and (195), we get

$$\begin{aligned} \frac{\partial E_n}{\partial a_k^\pi} &= -\sum_{j=1}^{K} \frac{\gamma_{nj}}{\pi_j} \pi_j (I_{jk} - \pi_k) \\ &= -\sum_{j=1}^{K} \gamma_{nj} (I_{jk} - \pi_k) = -\gamma_{nk} + \sum_{j=1}^{K} \gamma_{nj} \pi_k = \pi_k - \gamma_{nk}, \end{aligned}$$

where we have used the fact that, by (5.154), $\sum_{j=1}^{K} \gamma_{nj} = 1$ for all n.

5.35 **NOTE**: See Solution 5.34.

From (5.152) we have

$$a_{kl}^\mu = \mu_{kl}$$

and thus

$$\frac{\partial E_n}{\partial a_{kl}^\mu} = \frac{\partial E_n}{\partial \mu_{kl}}.$$

From (2.43), (5.153) and (5.154), we get

$$\begin{aligned} \frac{\partial E_n}{\partial \mu_{kl}} &= -\frac{\pi_k \mathcal{N}_{nk}}{\sum_{k'} \pi_{k'} \mathcal{N}_{nk'}} \frac{t_{nl} - \mu_{kl}}{\sigma_k^2(\mathbf{x}_n)} \\ &= \gamma_{nk} \left(\mathbf{t}_n | \mathbf{x}_n\right) \frac{\mu_{kl} - t_{nl}}{\sigma_k^2(\mathbf{x}_n)}. \end{aligned}$$

5.36 **NOTE**: In the 1st printing of PRML, equation (5.157) is incorrect and the correct equation appears at the end of this solution ; see also Solution 5.34.

From (5.151) and (5.153), we see that

$$\frac{\partial E_n}{\partial a_k^\sigma} = \frac{\partial E_n}{\partial \sigma_k} \frac{\partial \sigma_k}{\partial a_k^\sigma}, \tag{196}$$

where, from (5.151),

$$\frac{\partial \sigma_k}{\partial a_k^\sigma} = \sigma_k. \tag{197}$$

From (2.43), (5.153) and (5.154), we get

$$\begin{aligned}
\frac{\partial E_n}{\partial \sigma_k} &= -\frac{1}{\sum_{k'} \mathcal{N}_{nk'}} \left(\frac{L}{2\pi}\right)^{L/2} \left\{ -\frac{L}{\sigma^{L+1}} \exp\left(-\frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{2\sigma_k^2}\right) \right. \\
&\qquad \left. + \frac{1}{\sigma^L} \exp\left(-\frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{2\sigma_k^2}\right) \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^3} \right\} \\
&= \gamma_{nk} \left(\frac{L}{\sigma_k} - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^3}\right).
\end{aligned}$$

Combining this with (196) and (197), we get

$$\frac{\partial E_n}{\partial a_k^\sigma} = \gamma_{nk} \left(L - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2}\right).$$

5.37 From (2.59) and (5.148) we have

$$\begin{aligned}
\mathbb{E}\left[\mathbf{t}|\mathbf{x}\right] &= \int \mathbf{t} p\left(\mathbf{t}|\mathbf{x}\right) \, \mathrm{d}\mathbf{t} \\
&= \int \mathbf{t} \sum_{k=1}^{K} \pi_k(\mathbf{x}) \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\right) \, \mathrm{d}\mathbf{t} \\
&= \sum_{k=1}^{K} \pi_k(\mathbf{x}) \int \mathbf{t} \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\right) \, \mathrm{d}\mathbf{t} \\
&= \sum_{k=1}^{K} \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}).
\end{aligned}$$

We now introduce the shorthand notation

$$\bar{\mathbf{t}}_k = \boldsymbol{\mu}_k(\mathbf{x}) \quad \text{and} \quad \bar{\mathbf{t}} = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \bar{\mathbf{t}}_k.$$

Using this together with (2.59), (2.62), (5.148) and (5.158), we get

$$\begin{aligned} s^2(\mathbf{x}) &= \mathbb{E}\left[\|\mathbf{t}-\mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\|^2|\mathbf{x}\right] = \int \|\mathbf{t}-\overline{\mathbf{t}}\|^2 p\left(\mathbf{t}|\mathbf{x}\right)\,\mathrm{d}\mathbf{t} \\ &= \int \left(\mathbf{t}^{\mathrm{T}}\mathbf{t}-\mathbf{t}^{\mathrm{T}}\overline{\mathbf{t}}-\overline{\mathbf{t}}^{\mathrm{T}}\mathbf{t}+\overline{\mathbf{t}}^{\mathrm{T}}\overline{\mathbf{t}}\right)\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}),\sigma_k^2(\mathbf{x})\right)\,\mathrm{d}\mathbf{t} \\ &= \sum_{k=1}^{K}\pi_k(\mathbf{x})\left\{\sigma_k^2+\overline{\mathbf{t}}_k^{\mathrm{T}}\overline{\mathbf{t}}_k-\overline{\mathbf{t}}_k^{\mathrm{T}}\overline{\mathbf{t}}-\overline{\mathbf{t}}^{\mathrm{T}}\overline{\mathbf{t}}_k+\overline{\mathbf{t}}^{\mathrm{T}}\overline{\mathbf{t}}\right\} \\ &= \sum_{k=1}^{K}\pi_k(\mathbf{x})\left\{\sigma_k^2+\|\overline{\mathbf{t}}_k-\overline{\mathbf{t}}\|^2\right\} \\ &= \sum_{k=1}^{K}\pi_k(\mathbf{x})\left\{\sigma_k^2+\left\|\boldsymbol{\mu}_k(\mathbf{x})-\sum_{l}^{K}\pi_l\boldsymbol{\mu}_l(\mathbf{x})\right\|^2\right\}. \end{aligned}$$

5.38 Making the following substitions from the r.h.s. of (5.167) and (5.171),

$$\mathbf{x}\Rightarrow\mathbf{w}\quad\boldsymbol{\mu}\Rightarrow\mathbf{w}_{\mathrm{MAP}}\quad\boldsymbol{\Lambda}^{-1}\Rightarrow\mathbf{A}^{-1}$$

$$\mathbf{y}\Rightarrow t\quad\mathbf{A}\Rightarrow\mathbf{g}^{\mathrm{T}}\quad\mathbf{b}\Rightarrow y(\mathbf{x},\mathbf{w}_{\mathrm{MAP}})-\mathbf{g}^{\mathrm{T}}\mathbf{w}_{\mathrm{MAP}}\quad\mathbf{L}^{-1}\Rightarrow\beta^{-1},$$

in (2.113) and (2.114), (2.115) becomes

$$\begin{aligned} p(t) &= \mathcal{N}\left(t|\mathbf{g}^{\mathrm{T}}\mathbf{w}_{\mathrm{MAP}}+y(\mathbf{x},\mathbf{w}_{\mathrm{MAP}})-\mathbf{g}^{\mathrm{T}}\mathbf{w}_{\mathrm{MAP}},\beta^{-1}+\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}\right) \\ &= \mathcal{N}\left(t|y(\mathbf{x},\mathbf{w}_{\mathrm{MAP}}),\sigma^2\right), \end{aligned}$$

where σ^2 is defined by (5.173).

5.39 Using (4.135), we can approximate (5.174) as

$$p(\mathcal{D}|\alpha,\beta)\simeq p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}},\beta)p(\mathbf{w}_{\mathrm{MAP}}|\alpha)\int\exp\left\{-\frac{1}{2}\left(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}}\right)^{\mathrm{T}}\mathbf{A}\left(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}}\right)\right\}\mathrm{d}\mathbf{w},$$

where $\mathbf{A}$ is given by (5.166), as $p(\mathcal{D}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)$ is proportional to $p(\mathbf{w}|\mathcal{D},\alpha,\beta)$. Using (4.135), (5.162) and (5.163), we can rewrite this as

$$p(\mathcal{D}|\alpha,\beta)\simeq\prod_{n}^{N}\mathcal{N}(t_n|y(\mathbf{x}_n,\mathbf{w}_{\mathrm{MAP}}),\beta^{-1})\mathcal{N}(\mathbf{w}_{\mathrm{MAP}}|\mathbf{0},\alpha^{-1}\mathbf{I})\frac{(2\pi)^{W/2}}{|\mathbf{A}|^{1/2}}.$$

Taking the logarithm of both sides and then using (2.42) and (2.43), we obtain the desired result.

5.40 For a K-class neural network, the likelihood function is given by

$$\prod_n \prod_k y_k(\mathbf{x}_n, \mathbf{w})^{t_{nk}}$$

and the corresponding error function is given by (5.24).

Again we would use a Laplace approximation for the posterior distribution over the weights, but the corresponding Hessian matrix, $\mathbf{H}$, in (5.166), would now be derived from (5.24). Similarly, (5.24), would replace the binary cross entropy error term in the regularized error function (5.184).

The predictive distribution for a new pattern would again have to be approximated, since the resulting marginalization cannot be done analytically. However, in contrast to the two-class problem, there is no obvious candidate for this approximation, although Gibbs (1997) discusses various alternatives.

5.41 **NOTE**: In PRML, the final "const" term in Equation (5.183) should be ommitted.

This solutions is similar to Solution 5.39, with the difference that the log-likelihood term is now given by (5.181). Again using (4.135), the corresponding approximation of the marginal likelihood becomes

$$p(\mathcal{D}|\alpha) \simeq p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}})p(\mathbf{w}_{\mathrm{MAP}}|\alpha) \int \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})\right) \mathrm{d}\mathbf{w}, \quad (198)$$

where now

$$\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\mathbf{w}) = \mathbf{H} + \alpha\mathbf{I}.$$

Performing the integral in (198) using (4.135) and then taking the logarithm on, we get (5.183).

Chapter 6 Kernel Methods

6.1 We first of all note that $J(\mathbf{a})$ depends on $\mathbf{a}$ only through the form $\mathbf{Ka}$. Since typically the number N of data points is greater than the number M of basis functions, the matrix $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ will be rank deficient. There will then be M eigenvectors of $\mathbf{K}$ having non-zero eigenvalues, and $N-M$ eigenvectors with eigenvalue zero. We can then decompose $\mathbf{a} = \mathbf{a}_{\|} + \mathbf{a}_{\perp}$ where $\mathbf{a}_{\|}^{\mathrm{T}}\mathbf{a}_{\perp} = 0$ and $\mathbf{K}\mathbf{a}_{\perp} = \mathbf{0}$. Thus the value of $\mathbf{a}_{\perp}$ is not determined by $J(\mathbf{a})$. We can remove the ambiguity by setting $\mathbf{a}_{\perp} = \mathbf{0}$, or equivalently by adding a regularizer term

$$\frac{\epsilon}{2}\mathbf{a}_{\perp}^{\mathrm{T}}\mathbf{a}_{\perp}$$

to $J(\mathbf{a})$ where ϵ is a small positive constant. Then $\mathbf{a} = \mathbf{a}_\parallel$ where $\mathbf{a}_\parallel$ lies in the span of $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ and hence can be written as a linear combination of the columns of $\mathbf{\Phi}$, so that in component notation

$$a_n = \sum_{i=1}^{M} u_i \phi_i(\mathbf{x}_n)$$

or equivalently in vector notation

$$\mathbf{a} = \mathbf{\Phi}\mathbf{u}. \tag{199}$$

Substituting (199) into (6.7) we obtain

$$\begin{aligned} J(\mathbf{u}) &= \frac{1}{2}\left(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}\mathbf{u} \\ &= \frac{1}{2}\left(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} \end{aligned} \tag{200}$$

Since the matrix $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ has full rank we can define an equivalent parametrization given by

$$\mathbf{w} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u}$$

and substituting this into (200) we recover the original regularized error function (6.2).

6.2 Starting with an initial weight vector $\mathbf{w} = \mathbf{0}$ the Perceptron learning algorithm increments $\mathbf{w}$ with vectors $t_n\boldsymbol{\phi}(\mathbf{x}_n)$ where n indexes a pattern which is misclassified by the current model. The resulting weight vector therefore comprises a linear combination of vectors of the form $t_n\boldsymbol{\phi}(\mathbf{x}_n)$ which we can represent in the form

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}(\mathbf{x}_n) \tag{201}$$

where α_n is an integer specifying the number of times that pattern n was used to update $\mathbf{w}$ during training. The corresponding predictions made by the trained Perceptron are therefore given by

$$\begin{aligned} y(\mathbf{x}) &= \mathrm{sign}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right) \\ &= \mathrm{sign}\left(\sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right) \\ &= \mathrm{sign}\left(\sum_{n=1}^{N} \alpha_n t_n k(\mathbf{x}_n, \mathbf{x})\right). \end{aligned}$$

Thus the predictive function of the Perceptron has been expressed purely in terms of the kernel function. The learning algorithm of the Perceptron can similarly be written as

$$\alpha_n \to \alpha_n + 1$$

for patterns which are misclassified, in other words patterns which satisfy

$$t_n \left(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right) \geqslant 0.$$

Using (201) together with $\alpha_n \geqslant 0$, this can be written in terms of the kernel function in the form

$$t_n \left(\sum_{m=1}^{N} k(\mathbf{x}_m, \mathbf{x}_n) \right) \geqslant 0$$

and so the learning algorithm depends only on the elements of the Gram matrix.

6.3 The distance criterion for the nearest neighbour classifier can be expressed in terms of the kernel as follows

$$\begin{aligned} D(\mathbf{x}, \mathbf{x}_n) &= \|\mathbf{x} - \mathbf{x}_n\|^2 \\ &= \mathbf{x}^{\mathrm{T}}\mathbf{x} + \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - 2\mathbf{x}^{\mathrm{T}}\mathbf{x}_n \\ &= k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}_n, \mathbf{x}_n) - 2k(\mathbf{x}, \mathbf{x}_n) \end{aligned}$$

where $k(\mathbf{x}, \mathbf{x}_n) = \mathbf{x}^{\mathrm{T}}\mathbf{x}_n$. We then obtain a non-linear kernel classifier by replacing the linear kernel with some other choice of kernel function.

6.4 An example of such a matrix is

$$\begin{pmatrix} 2 & -2 \\ -3 & 4 \end{pmatrix}.$$

We can verify this by calculating the determinant of

$$\begin{pmatrix} 2-\lambda & -2 \\ -3 & 4-\lambda \end{pmatrix},$$

setting the resulting expression equal to zero and solve for the eigenvalues λ, yielding

$$\lambda_1 \simeq 5.65 \quad \text{and} \quad \lambda_2 \simeq 0.35,$$

which are both positive.

6.5 The results (6.13) and (6.14) are easily proved by using (6.1) which defines the kernel in terms of the scalar product between the feature vectors for two input vectors. If $k_1(\mathbf{x}, \mathbf{x}')$ is a valid kernel then there must exist a feature vector $\boldsymbol{\phi}(\mathbf{x})$ such that

$$k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}').$$

It follows that

$$ck_1(\mathbf{x}, \mathbf{x}') = \mathbf{u}(\mathbf{x})^{\mathrm{T}} \mathbf{u}(\mathbf{x}')$$

where

$$\mathbf{u}(\mathbf{x}) = c^{1/2} \boldsymbol{\phi}(\mathbf{x})$$

and so $ck_1(\mathbf{x}, \mathbf{x}')$ can be expressed as the scalar product of feature vectors, and hence is a valid kernel.

Similarly, for (6.14) we can write

$$f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') = \mathbf{v}(\mathbf{x})^{\mathrm{T}}\mathbf{v}(\mathbf{x}')$$

where we have defined

$$\mathbf{v}(\mathbf{x}) = f(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}).$$

Again, we see that $f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')$ can be expressed as the scalar product of feature vectors, and hence is a valid kernel.

Alternatively, these results can be proved be appealing to the general result that the Gram matrix, $\mathbf{K}$, whose elements are given by $k(\mathbf{x}_n, \mathbf{x}_m)$, should be positive semidefinite for all possible choices of the set $\{\mathbf{x}_n\}$, by following a similar argument to Solution 6.7 below.

6.6 Equation (6.15) follows from (6.13), (6.17) and (6.18).

For (6.16), we express the exponential as a power series, yielding

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \exp(k_1(\mathbf{x}, \mathbf{x}')) \\ &= \sum_{m=0}^{\infty} \frac{(k_1(\mathbf{x}, \mathbf{x}'))^m}{m!}. \end{aligned}$$

Since this is a polynomial in $k_1(\mathbf{x}, \mathbf{x}')$ with positive coefficients, (6.16) follows from (6.15).

6.7 (6.17) is most easily proved by making use of the result, discussed on page 295, that a necessary and sufficient condition for a function $k(\mathbf{x}, \mathbf{x}')$ to be a valid kernel is that the Gram matrix $\mathbf{K}$, whose elements are given by $k(\mathbf{x}_n, \mathbf{x}_m)$, should be positive semidefinite for all possible choices of the set $\{\mathbf{x}_n\}$. A matrix $\mathbf{K}$ is positive semi-definite if, and only if,

$$\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a} \geqslant 0$$

for any choice of the vector $\mathbf{a}$. Let $\mathbf{K}_1$ be the Gram matrix for $k_1(\mathbf{x}, \mathbf{x}')$ and let $\mathbf{K}_2$ be the Gram matrix for $k_2(\mathbf{x}, \mathbf{x}')$. Then

$$\mathbf{a}^{\mathrm{T}}(\mathbf{K}_1 + \mathbf{K}_2)\mathbf{a} = \mathbf{a}^{\mathrm{T}}\mathbf{K}_1\mathbf{a} + \mathbf{a}^{\mathrm{T}}\mathbf{K}_2\mathbf{a} \geqslant 0$$

where we have used the fact that $\mathbf{K}_1$ and $\mathbf{K}_2$ are positive semi-definite matrices, together with the fact that the sum of two non-negative numbers will itself be non-negative. Thus, (6.17) defines a valid kernel.

To prove (6.18), we take the approach adopted in Solution 6.5. Since we know that $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$ are valid kernels, we know that there exist mappings $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\psi}(\mathbf{x})$ such that

$$k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}') \qquad \text{and} \qquad k_2(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}').$$

Hence

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') \\ &= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')\,\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}') \\ &= \sum_{m=1}^{M}\phi_m(\mathbf{x})\phi_m(\mathbf{x}')\sum_{n=1}^{N}\psi_n(\mathbf{x})\psi_n(\mathbf{x}') \\ &= \sum_{m=1}^{M}\sum_{n=1}^{N}\phi_m(\mathbf{x})\phi_m(\mathbf{x}')\psi_n(\mathbf{x})\psi_n(\mathbf{x}') \\ &= \sum_{k=1}^{K}\varphi_k(\mathbf{x})\varphi_k(\mathbf{x}') \\ &= \boldsymbol{\varphi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}'), \end{aligned}$$

where $K = MN$ and

$$\varphi_k(\mathbf{x}) = \phi_{((k-1)\oslash N)+1}(\mathbf{x})\psi_{((k-1)\odot N)+1}(\mathbf{x}),$$

where in turn $\oslash$ and $\odot$ denote integer division and remainder, respectively.

6.8 If we consider the Gram matrix, $\mathbf{K}$, corresponding to the l.h.s. of (6.19), we have

$$(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = k_3\left(\boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j)\right) = (\mathbf{K}_3)_{ij}$$

where $\mathbf{K}_3$ is the Gram matrix corresponding to $k_3(\cdot,\cdot)$. Since $k_3(\cdot,\cdot)$ is a valid kernel,

$$\mathbf{u}^{\mathrm{T}}\mathbf{K}\mathbf{u} = \mathbf{u}^{\mathrm{T}}\mathbf{K}_3\mathbf{u} \geqslant 0.$$

For (6.20), let $\mathbf{K} = \mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}$, so that $(\mathbf{K})_{ij} = \mathbf{x}_i^{\mathrm{T}}\mathbf{A}\mathbf{x}_j$, and consider

$$\begin{aligned} \mathbf{u}^{\mathrm{T}}\mathbf{K}\mathbf{u} &= \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}\mathbf{u} \\ &= \mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v} \geqslant 0 \end{aligned}$$

where , $\mathbf{v} = \mathbf{X}\mathbf{u}$ and we have used that $\mathbf{A}$ is positive semidefinite.

6.9 Equations (6.21) and (6.22) are special cases of (6.17) and (6.18), respectively, where $k_a(\cdot,\cdot)$ and $k_b(\cdot,\cdot)$ only depend on particular elements in their argument vectors. Thus (6.21) and (6.22) follow from the more general results.

6.10 Any solution of a linear learning machine based on this kernel must take the form

$$y(\mathbf{x}) = \sum_{n=1}^{N}\alpha_n k(\mathbf{x}_n, \mathbf{x}) = \left(\sum_{n=1}^{N}\alpha_n f(\mathbf{x}_n)\right) f(\mathbf{x}) = C f(\mathbf{x}).$$

6.11 As discussed in Solution 6.6, the exponential kernel (6.16) can be written as an infinite sum of terms, each of which can itself be written as an inner product of feature vectors, according to (6.15). Thus, by concatenating the feature vectors of the indvidual terms in that sum, we can write this as an inner product of infinite dimension feature vectors. More formally,

$$\begin{aligned}\exp\left(\mathbf{x}^{\mathrm{T}}\mathbf{x}'/\sigma^2\right) &= \sum_{m=0}^{\infty}\boldsymbol{\phi}_m(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}_0(\mathbf{x}') \\ &= \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}')\end{aligned}$$

where $\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}} = \left[\boldsymbol{\phi}_0(\mathbf{x})^{\mathrm{T}}, \boldsymbol{\phi}_1(\mathbf{x})^{\mathrm{T}}, \ldots\right]$. Hence, we can write (6.23) as

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\varphi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}')$$

where

$$\boldsymbol{\varphi}(\mathbf{x}) = \exp\left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}}{\sigma^2}\right)\boldsymbol{\psi}(\mathbf{x}).$$

6.12 **NOTE**: In the 1st printing of PRML, there is an error in the text relating to this exercise. Immediately following (6.27), it says: $|A|$ denotes the number of *subsets* in A; it should have said: $|A|$ denotes the number of *elements* in A.

Since A may be equal to D (the subset relation was not defined to be strict), $\boldsymbol{\phi}(D)$ must be defined. This will map to a vector of $2^{|D|}$ 1s, one for each possible subset of D, including D itself as well as the empty set. For $A \subset D$, $\boldsymbol{\phi}(A)$ will have 1s in all positions that correspond to subsets of A and 0s in all other positions. Therefore, $\boldsymbol{\phi}(A_1)^{\mathrm{T}}\boldsymbol{\phi}(A_2)$ will count the number of subsets shared by A_1 and A_2. However, this can just as well be obtained by counting the number of elements in the intersection of A_1 and A_2, and then raising 2 to this number, which is exactly what (6.27) does.

6.13 In the case of the transformed parameter $\boldsymbol{\psi}(\boldsymbol{\theta})$, we have

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{M}\mathbf{g}_{\boldsymbol{\psi}} \tag{202}$$

where $\mathbf{M}$ is a matrix with elements

$$M_{ij} = \frac{\partial\psi_i}{\partial\theta_j}$$

(recall that $\boldsymbol{\psi}(\boldsymbol{\theta})$ is assumed to be differentiable) and

$$\mathbf{g}_{\boldsymbol{\psi}} = \nabla_{\boldsymbol{\psi}} \ln p\left(\mathbf{x}|\boldsymbol{\psi}(\boldsymbol{\theta})\right).$$

The Fisher information matrix then becomes

$$\begin{aligned}\mathbf{F} &= \mathbb{E}_{\mathbf{x}}\left[\mathbf{M}\mathbf{g}_{\boldsymbol{\psi}}\mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\right] \\ &= \mathbf{M}\mathbb{E}_{\mathbf{x}}\left[\mathbf{g}_{\boldsymbol{\psi}}\mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}}\right]\mathbf{M}^{\mathrm{T}}.\end{aligned} \tag{203}$$

Substituting (202) and (203) into (6.33), we get

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \mathbf{M}^{\mathrm{T}} \left(\mathbf{M} \mathbb{E}_{\mathbf{x}} \left[\mathbf{g}_{\boldsymbol{\psi}} \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \right] \mathbf{M}^{\mathrm{T}} \right)^{-1} \mathbf{M} \mathbf{g}_{\boldsymbol{\psi}} \\ &= \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \mathbf{M}^{\mathrm{T}} \left(\mathbf{M}^{\mathrm{T}} \right)^{-1} \mathbb{E}_{\mathbf{x}} \left[\mathbf{g}_{\boldsymbol{\psi}} \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \right]^{-1} \mathbf{M}^{-1} \mathbf{M} \mathbf{g}_{\boldsymbol{\psi}} \\ &= \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \mathbb{E}_{\mathbf{x}} \left[\mathbf{g}_{\boldsymbol{\psi}} \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \right]^{-1} \mathbf{g}_{\boldsymbol{\psi}}, \end{aligned} \tag{204}$$

where we have used (C.3) and the fact that $\boldsymbol{\psi}(\boldsymbol{\theta})$ is assumed to be invertible. Since $\boldsymbol{\theta}$ was simply replaced by $\boldsymbol{\psi}(\boldsymbol{\theta})$, (204) corresponds to the original form of (6.33).

6.14 In order to evaluate the Fisher kernel for the Gaussian we first note that the covariance is assumed to be fixed, and hence the parameters comprise only the elements of the mean $\boldsymbol{\mu}$. The first step is to evaluate the Fisher score defined by (6.32). From the definition (2.43) of the Gaussian we have

$$\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) = \nabla_{\boldsymbol{\mu}} \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}) = \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

Next we evaluate the Fisher information matrix using the definition (6.34), giving

$$\mathbf{F} = \mathbb{E}_{\mathbf{x}} \left[\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) \mathbf{g}(\boldsymbol{\mu}, \mathbf{x})^{\mathrm{T}} \right] = \mathbf{S}^{-1} \mathbb{E}_{\mathbf{x}} \left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \right] \mathbf{S}^{-1}.$$

Here the expectation is with respect to the original Gaussian distribution, and so we can use the standard result

$$\mathbb{E}_{\mathbf{x}} \left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \right] = \mathbf{S}$$

from which we obtain

$$\mathbf{F} = \mathbf{S}^{-1}.$$

Thus the Fisher kernel is given by

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}' - \boldsymbol{\mu}),$$

which we note is just the squared Mahalanobis distance.

6.15 The determinant for the 2×2 Gram matrix

$$\begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{pmatrix}$$

equals

$$k(x_1, x_1) k(x_2, x_2) - k(x_1, x_2)^2,$$

where we have used the fact that $k(x_1, x_2) = k(x_2, x_1)$. Then (6.96) follows directly from the fact that this determinant must be non-negative for a positive semidefinite matrix.

6.16 **NOTE**: In the 1^{st} printing of PRML, a detail is missing in this exercise; the text "where $\mathbf{w}_{\perp}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) = 0$ for all n," should be inserted at the beginning of the line immediately following equation (6.98).

We start by rewriting (6.98) as

$$\mathbf{w} = \mathbf{w}_{\|} + \mathbf{w}_{\perp} \tag{205}$$

where

$$\mathbf{w}_{\|} = \sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(\mathbf{x}_n).$$

Note that since $\mathbf{w}_{\perp}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) = 0$ for all n,

$$\mathbf{w}_{\perp}^{\mathrm{T}}\mathbf{w}_{\|} = 0. \tag{206}$$

Using (205) and (206) together with the fact that $\mathbf{w}_{\perp}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) = 0$ for all n, we can rewrite (6.97) as

$$\begin{aligned} J(\mathbf{w}) &= f\left((\mathbf{w}_{\|} + \mathbf{w}_{\perp})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_1), \ldots, (\mathbf{w}_{\|} + \mathbf{w}_{\perp})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_N)\right) \\ &\quad + g\left((\mathbf{w}_{\|} + \mathbf{w}_{\perp})^{\mathrm{T}}(\mathbf{w}_{\|} + \mathbf{w}_{\perp})\right) \\ &= f\left(\mathbf{w}_{\|}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \mathbf{w}_{\|}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_N)\right) + g\left(\mathbf{w}_{\|}^{\mathrm{T}}\mathbf{w}_{\|} + \mathbf{w}_{\perp}^{\mathrm{T}}\mathbf{w}_{\perp}\right). \end{aligned}$$

Since $g(\cdot)$ is monotonically increasing, it will have its minimum w.r.t. $\mathbf{w}_{\perp}$ at $\mathbf{w}_{\perp} = \mathbf{0}$, in which case

$$\mathbf{w} = \mathbf{w}_{\|} = \sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(\mathbf{x}_n)$$

as desired.

6.17 **NOTE**: In the 1st printing of PRML, there are typographical errors in the text relating to this exercise. In the sentence following immediately after (6.39), $f(\mathbf{x})$ should be replaced by $y(\mathbf{x})$. Also, on the l.h.s. of (6.40), $y(\mathbf{x}_n)$ should be replaced by $y(\mathbf{x})$. There were also errors in Appendix D, which might cause confusion; please consult the errata on the PRML website.

Following the discussion in Appendix D we give a first-principles derivation of the solution. First consider a variation in the function $y(\mathbf{x})$ of the form

$$y(\mathbf{x}) \to y(\mathbf{x}) + \epsilon\eta(\mathbf{x}).$$

Substituting into (6.39) we obtain

$$E[y + \epsilon\eta] = \frac{1}{2}\sum_{n=1}^{N}\int \{y(\mathbf{x}_n + \boldsymbol{\xi}) + \epsilon\eta(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\}^2 \nu(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}.$$

Now we expand in powers of ϵ and set the coefficient of ϵ, which corresponds to the functional first derivative, equal to zero, giving

$$\sum_{n=1}^{N}\int \{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\}\,\eta(\mathbf{x}_n + \boldsymbol{\xi})\nu(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = 0. \tag{207}$$

This must hold for every choice of the variation function $\eta(\mathbf{x})$. Thus we can choose

$$\eta(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{z})$$

where $\delta(\,\cdot\,)$ is the Dirac delta function. This allows us to evaluate the integral over $\boldsymbol{\xi}$ giving

$$\sum_{n=1}^{N} \int \{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\}\, \delta(\mathbf{x}_n + \boldsymbol{\xi} - \mathbf{z}) \nu(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi} = \sum_{n=1}^{N} \{y(\mathbf{z}) - t_n\}\, \nu(\mathbf{z} - \mathbf{x}_n).$$

Substituting this back into (207) and rearranging we then obtain the required result (6.40).

6.18 From the product rule we have

$$p(t|x) = \frac{p(t, x)}{p(x)}.$$

With $p(t, x)$ given by (6.42) and

$$f(x - x_n, t - t_n) = \mathcal{N}\left([x - x_n, t - t_n]^{\mathrm{T}} | \mathbf{0}, \sigma^2 \mathbf{I}\right)$$

this becomes

$$\begin{aligned} p(t|x) &= \frac{\sum_{n=1}^{N} \mathcal{N}\left([x - x_n, t - t_n]^{\mathrm{T}} | \mathbf{0}, \sigma^2 \mathbf{I}\right)}{\int \sum_{m=1}^{N} \mathcal{N}\left([x - x_m, t - t_m]^{\mathrm{T}} | \mathbf{0}, \sigma^2 \mathbf{I}\right)\, \mathrm{d}t} \\ &= \frac{\sum_{n=1}^{N} \mathcal{N}\left(x - x_n | 0, \sigma^2\right) \mathcal{N}\left(t - t_n | 0, \sigma^2\right)}{\sum_{m=1}^{N} \mathcal{N}\left(x - x_m | 0, \sigma^2\right)}. \end{aligned}$$

From (6.46), (6.47), the definition of $f(x, t)$ and the properties of the Gaussian distribution, we can rewrite this as

$$\begin{aligned} p(t|x) &= \sum_{n=1}^{N} k(x, x_n) \mathcal{N}\left(t - t_n | 0, \sigma^2\right) \\ &= \sum_{n=1}^{N} k(x, x_n) \mathcal{N}\left(t | t_n, \sigma^2\right) \end{aligned} \tag{208}$$

where

$$k(x, x_n) = \frac{\mathcal{N}\left(x - x_n | 0, \sigma^2\right)}{\sum_{m=1}^{N} \mathcal{N}\left(x - x_m | 0, \sigma^2\right)}.$$

We see that this a Gaussian mixture model where $k(x, x_n)$ play the role of input dependent mixing coefficients.

Using (208) it is straightforward to calculate various expectations:

$$\begin{aligned}\mathbb{E}[t|x] &= \int t\, p(t|x)\, \mathrm{d}t \\ &= \int t \sum_{n=1}^{N} k(x, x_n) \mathcal{N}\left(t|t_n, \sigma^2\right)\, \mathrm{d}t \\ &= \sum_{n=1}^{N} k(x, x_n) \int t\, \mathcal{N}\left(t|t_n, \sigma^2\right)\, \mathrm{d}t \\ &= \sum_{n=1}^{N} k(x, x_n)\, t_n\end{aligned}$$

and

$$\begin{aligned}\mathrm{var}[t|x] &= \mathbb{E}\left[(t - \mathbb{E}[t|x])^2\right] \\ &= \int (t - \mathbb{E}[t|x])^2\, p(t|x)\, \mathrm{d}t \\ &= \sum_{n=1}^{N} k(x, x_n) \int (t - \mathbb{E}[t|x])^2\, \mathcal{N}\left(t|t_n, \sigma^2\right)\, \mathrm{d}t \\ &= \sum_{n=1}^{N} k(x, x_n) \left(\sigma^2 + t_n^2 - 2t_n \mathbb{E}[t|x] + \mathbb{E}[t|x]^2\right) \\ &= \sigma^2 - \mathbb{E}[t|x]^2 + \sum_{n=1}^{N} k(x, x_n)\, t_n^2.\end{aligned}$$

6.19 Changing variables to $\mathbf{z}_n = \mathbf{x}_n - \boldsymbol{\xi}_n$ we obtain

$$E = \frac{1}{2} \sum_{n=1}^{N} \int [y(\mathbf{z}_n) - t_n]^2\, g(\mathbf{x}_n - \mathbf{z}_n)\, \mathrm{d}\mathbf{z}_n.$$

If we set the functional derivative of E with respect to the function $y(\mathbf{x})$, for some general value of $\mathbf{x}$, to zero using the calculus of variations (see Appendix D) we have

$$\begin{aligned}\frac{\delta E}{\delta \mathbf{y}(\mathbf{x})} &= \sum_{n=1}^{N} \int [y(\mathbf{z}_n) - t_n]\, g(\mathbf{x}_n - \mathbf{z}_n) \delta(\mathbf{x} - \mathbf{z}_n)\, \mathrm{d}\mathbf{z}_n \\ &= \sum_{n=1}^{N} [y(\mathbf{x}) - t_n]\, g(\mathbf{x}_n - \mathbf{x}) = 0.\end{aligned}$$

Solving for $y(\mathbf{x})$ we obtain

$$y(\mathbf{x}) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n \tag{209}$$

where we have defined

$$k(\mathbf{x}, \mathbf{x}_n) = \frac{g(\mathbf{x}_n - \mathbf{x})}{\sum_n g(\mathbf{x}_n - \mathbf{x})}.$$

This an expansion in kernel functions, where the kernels satisfy the summation constraint $\sum_n k(\mathbf{x}, \mathbf{x}_n) = 1$.

6.20 Given the joint distribution (6.64), we can identify t_{N+1} with $\mathbf{x}_a$ and $\mathbf{t}$ with $\mathbf{x}_b$ in (2.65). Note that this means that we are prepending rather than appending t_{N+1} to $\mathbf{t}$ and $\mathbf{C}_{N+1}$ therefore gets redefined as

$$\mathbf{C}_{N+1} = \begin{pmatrix} c & \mathbf{k}^{\mathrm{T}} \\ \mathbf{k} & \mathbf{C}_N \end{pmatrix}.$$

It then follows that

$$\begin{array}{lll} \boldsymbol{\mu}_a = 0 & \boldsymbol{\mu}_b = \mathbf{0} & \mathbf{x}_b = \mathbf{t} \\ \boldsymbol{\Sigma}_{aa} = c & \boldsymbol{\Sigma}_{bb} = \mathbf{C}_N & \boldsymbol{\Sigma}_{ab} = \boldsymbol{\Sigma}_{ba}^{\mathrm{T}} = \mathbf{k}^{\mathrm{T}} \end{array}$$

in (2.81) and (2.82), from which (6.66) and (6.67) follows directly.

6.21 Both the Gaussian process and the linear regression model give rise to Gaussian predictive distributions $p(t_{N+1}|\mathbf{x}_{N+1})$ so we simply need to show that these have the same mean and variance. To do this we make use of the expression (6.54) for the kernel function defined in terms of the basis functions. Using (6.62) the covariance matrix $\mathbf{C}_N$ then takes the form

$$\mathbf{C}_N = \frac{1}{\alpha}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N \tag{210}$$

where $\boldsymbol{\Phi}$ is the design matrix with elements $\Phi_{nk} = \phi_k(\mathbf{x}_n)$, and $\mathbf{I}_N$ denotes the $N \times N$ unit matrix. Consider first the mean of the Gaussian process predictive distribution, which from (210), (6.54), (6.66) and the definitions in the text preceding (6.66) is given by

$$m_{N+1} = \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\left(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\mathbf{t}.$$

We now make use of the matrix identity (C.6) to give

$$\boldsymbol{\Phi}^{\mathrm{T}}\left(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1} = \alpha\beta\left(\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \alpha\mathbf{I}_M\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}} = \alpha\beta\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}.$$

Thus the mean becomes

$$m_{N+1} = \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

which we recognize as the mean of the predictive distribution for the linear regression model given by (3.58) with $\mathbf{m}_N$ defined by (3.53) and $\mathbf{S}_N$ defined by (3.54).

For the variance we similarly substitute the expression (210) for the kernel function into the Gaussian process variance given by (6.67) and then use (6.54) and the definitions in the text preceding (6.66) to obtain

$$\begin{aligned}\sigma^2_{N+1}(\mathbf{x}_{N+1}) &= \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_{N+1}) + \beta^{-1} \\ &\quad -\alpha^{-2}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\left(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}_{N+1}) \\ &= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\big(\alpha^{-1}\mathbf{I}_M \\ &\quad -\alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\left(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\big)\boldsymbol{\phi}(\mathbf{x}_{N+1}). \end{aligned} \tag{211}$$

We now make use of the matrix identity (C.7) to give

$$\begin{aligned}&\alpha^{-1}\mathbf{I}_M - \alpha^{-1}\mathbf{I}_M\boldsymbol{\Phi}^{\mathrm{T}}\left(\boldsymbol{\Phi}(\alpha^{-1}\mathbf{I}_M)\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\alpha^{-1}\mathbf{I}_M \\ &\qquad = \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1} = \mathbf{S}_N,\end{aligned}$$

where we have also used (3.54). Substituting this in (211), we obtain

$$\sigma^2_N(\mathbf{x}_{N+1}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})$$

as derived for the linear regression model in Section 3.3.2.

6.22 From (6.61) we have

$$p\left(\begin{bmatrix}\mathbf{t}_{1\ldots N} \\ \mathbf{t}_{N+1\ldots N+L}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{t}_{1\ldots N} \\ \mathbf{t}_{N+1\ldots N+L}\end{bmatrix}\,\middle|\,\mathbf{0}, \mathbf{C}\right)$$

with $\mathbf{C}$ specified by (6.62).

For our purposes, it is useful to consider the following partition[2] of $\mathbf{C}$:

$$\mathbf{C} = \begin{pmatrix}\mathbf{C}_{bb} & \mathbf{C}_{ba} \\ \mathbf{C}_{ab} & \mathbf{C}_{aa}\end{pmatrix},$$

where $\mathbf{C}_{aa}$ corresponds to $\mathbf{t}_{N+1\ldots N+L}$ and $\mathbf{C}_{bb}$ corresponds to $\mathbf{t}_{1\ldots N}$. We can use this together with (2.94)–(2.97) and (6.61) to obtain the conditional distribution

$$p(\mathbf{t}_{N+1\ldots N+L}|\mathbf{t}_{1\ldots N}) = \mathcal{N}\left(\mathbf{t}_{N+1\ldots N+L}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}^{-1}\right) \tag{212}$$

where, from (2.78)–(2.80),

$$\begin{aligned}\boldsymbol{\Lambda}^{-1}_{aa} &= \mathbf{C}_{aa} - \mathbf{C}_{ab}\mathbf{C}^{-1}_{bb}\mathbf{C}_{ba} \\ \boldsymbol{\Lambda}_{ab} &= -\boldsymbol{\Lambda}_{aa}\mathbf{C}_{ab}\mathbf{C}^{-1}_{bb}\end{aligned} \tag{213}$$

[2]The indexing and ordering of this partition have been chosen to match the indexing used in (2.94)–(2.97) as well as the ordering of elements used in the single variate case, as seen in (6.64)–(6.65).

and

$$\boldsymbol{\mu}_{a|b} = -\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\mathbf{t}_{1\ldots N} = \mathbf{C}_{ab}\mathbf{C}_{bb}^{-1}\mathbf{t}_{1\ldots N}. \tag{214}$$

Restricting (212) to a single test target, we obtain the corresponding marginal distribution, where $\mathbf{C}_{aa}$, $\mathbf{C}_{ba}$ and $\mathbf{C}_{bb}$ correspond to c, $\mathbf{k}$ and $\mathbf{C}_N$ in (6.65), respectively. Making the matching substitutions in (213) and (214), we see that they equal (6.67) and (6.66), respectively.

6.23 **NOTE**: In the 1st printing of PRML, a typographical mistake appears in the text of the exercise at line three, where it should say "... a training set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$".

If we assume that the target variables, $t_1, \ldots, t_D$, are independent given the input vector, $\mathbf{x}$, this extension is straightforward.

Using analogous notation to the univariate case,

$$p(\mathbf{t}_{N+1}|\mathbf{T}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{m}(\mathbf{x}_{N+1}), \sigma(\mathbf{x}_{N+1})\mathbf{I}),$$

where $\mathbf{T}$ is a $N \times D$ matrix with the vectors $\mathbf{t}_1^{\mathrm{T}}, \ldots, \mathbf{t}_N^{\mathrm{T}}$ as its rows,

$$\mathbf{m}(\mathbf{x}_{N+1})^{\mathrm{T}} = \mathbf{k}^{\mathrm{T}}\mathbf{C}_N\mathbf{T}$$

and $\sigma(\mathbf{x}_{N+1})$ is given by (6.67). Note that $\mathbf{C}_N$, which only depend on the input vectors, is the same in the uni- and multivariate models.

6.24 Since the diagonal elements of a diagonal matrix are also the eigenvalues of the matrix, $\mathbf{W}$ is positive definite (see Appendix C). Alternatively, for an arbitrary, non-zero vector $\mathbf{x}$,

$$\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} = \sum_i x_i^2 W_{ii} > 0.$$

If $\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} > 0$ and $\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x} > 0$ for an arbitrary, non-zero vector $\mathbf{x}$, then

$$\mathbf{x}^{\mathrm{T}}(\mathbf{W} + \mathbf{V})\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x} > 0.$$

6.25 Substituting the gradient and the Hessian into the Newton-Raphson formula we obtain

$$\begin{aligned}
\mathbf{a}_N^{\mathrm{new}} &= \mathbf{a}_N + (\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1}\mathbf{a}_N\right] \\
&= (\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N\mathbf{a}_N\right] \\
&= \mathbf{C}_N(\mathbf{I} + \mathbf{W}_N\mathbf{C}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N\mathbf{a}_N\right]
\end{aligned}$$

6.26 Using (2.115) the mean of the posterior distribution $p(a_{N+1}|\mathbf{t}_N)$ is given by

$$\mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{a}_N^*.$$

Combining this with the condition

$$\mathbf{C}_N^{-1}\mathbf{a}_N^* = \mathbf{t}_N - \boldsymbol{\sigma}_N$$

satisfied by $\mathbf{a}_N^*$ we obtain (6.87).

Similarly, from (2.115) the variance of the posterior distribution $p(a_{N+1}|\mathbf{t}_N)$ is given by

$$\begin{aligned}\mathrm{var}[a_{N+1}|\mathbf{t}_N] &= c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k} + \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{C}_N(\mathbf{I}+\mathbf{W}_N\mathbf{C}_N)^{-1}\mathbf{C}_N^{-1}\mathbf{k}\\ &= c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\left[\mathbf{I} - (\mathbf{C}_N^{-1}+\mathbf{W}_N)^{-1}\mathbf{C}_N^{-1}\right]\mathbf{k}\\ &= c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}(\mathbf{C}_N^{-1}+\mathbf{W}_N)^{-1}\mathbf{W}_N\mathbf{k}\\ &= c - \mathbf{k}^{\mathrm{T}}(\mathbf{W}_N^{-1}+\mathbf{C}_N)^{-1}\mathbf{k}\end{aligned}$$

as required.

6.27 Using (4.135), (6.80) and (6.85), we can approximate (6.89) as follows:

$$\begin{aligned}p(\mathbf{t}_N|\boldsymbol{\theta}) &= \int p(\mathbf{t}_N|\mathbf{a}_N)p(\mathbf{a}_N|\boldsymbol{\theta})\,\mathrm{d}\mathbf{a}_N\\ &\simeq p(\mathbf{t}_N|\mathbf{a}_N^\star)p(\mathbf{a}_N^\star|\boldsymbol{\theta})\\ &\qquad \int \exp\left\{-\frac{1}{2}\left(\mathbf{a}_N-\mathbf{a}_N^\star\right)^{\mathrm{T}}\mathbf{H}\left(\mathbf{a}_N-\mathbf{a}_N^\star\right)\right\}\mathrm{d}\mathbf{a}_N\\ &= \exp\left(\Psi(\mathbf{a}_N^\star)\right)\frac{(2\pi)^{N/2}}{|\mathbf{H}|^{1/2}}.\end{aligned}$$

Taking the logarithm, we obtain (6.90).

To derive (6.91), we gather the terms from (6.90) that involve $\mathbf{C}_N$, yielding

$$\begin{aligned}&-\frac{1}{2}\left(\mathbf{a}_N^{\star\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{a}_N^\star + \ln|\mathbf{C}_N| + \ln|\mathbf{W}_N+\mathbf{C}_N^{-1}|\right)\\ &\qquad = -\frac{1}{2}\mathbf{a}_N^{\star\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{a}_N^\star - \frac{1}{2}\ln|\mathbf{C}_N\mathbf{W}_N+\mathbf{I}|.\end{aligned}$$

Applying (C.21) and (C.22) to the first and second terms, respectively, we get (6.91).

Applying (C.22) to the l.h.s. of (6.92), we get

$$\begin{aligned}-\frac{1}{2}\sum_{n=1}^{N}\frac{\partial\ln|\mathbf{W}_N+\mathbf{C}_N^{-1}|}{\partial a_n^\star}\frac{\partial a_n^\star}{\partial\theta_j} &= -\frac{1}{2}\sum_{n=1}^{N}\mathrm{Tr}\left(\left(\mathbf{W}_N+\mathbf{C}_N^{-1}\right)^{-1}\frac{\partial\mathbf{W}}{\partial a_n^\star}\right)\frac{\partial a_n^\star}{\partial\theta_j}\\ &= -\frac{1}{2}\sum_{n=1}^{N}\mathrm{Tr}\left(\left(\mathbf{C}_N\mathbf{W}_N+\mathbf{I}\right)^{-1}\mathbf{C}_N\frac{\partial\mathbf{W}}{\partial a_n^\star}\right)\frac{\partial a_n^\star}{\partial\theta_j}. \quad (215)\end{aligned}$$

Using the definition of $\mathbf{W}$ together with (4.88), we have

$$\begin{aligned}\frac{\mathrm{d}W_{nn}}{\mathrm{d}a_n^\star} &= \frac{\mathrm{d}\sigma_n^\star(1-\sigma_n^\star)}{\mathrm{d}a_n^\star}\\ &= \sigma_n^\star(1-\sigma_n^\star)^2 - \sigma_n^{\star 2}(1-\sigma_n^\star)\\ &= \sigma_n^\star(1-\sigma_n^\star)(1-2\sigma_n^\star)\end{aligned}$$

and substituting this into (215) we the the r.h.s. of (6.92).

Gathering all the terms in (6.93) involving $\partial a_n^\star / \partial \theta_j$ on one side, we get

$$(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N) \frac{\partial a_n^\star}{\partial \theta_j} = \frac{\partial \mathbf{C}_N}{\partial \theta_j} (\mathbf{t}_N - \boldsymbol{\sigma}_N).$$

Left-multiplying both sides with $(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N)^{-1}$, we obtain (6.94).

Chapter 7 Sparse Kernel Machines

7.1 From Bayes' theorem we have

$$p(t|\mathbf{x}) \propto p(\mathbf{x}|t)p(t)$$

where, from (2.249),

$$p(\mathbf{x}|t) = \frac{1}{N_t} \sum_{n=1}^{N} \frac{1}{Z_k} k(\mathbf{x}, \mathbf{x}_n) \delta(t, t_n).$$

Here N_t is the number of input vectors with label t ($+1$ or -1) and $N = N_{+1} + N_{-1}$. $\delta(t, t_n)$ equals 1 if $t = t_n$ and 0 otherwise. Z_k is the normalisation constant for the kernel. The minimum misclassification-rate is achieved if, for each new input vector, $\tilde{\mathbf{x}}$, we chose $\tilde{t}$ to maximise $p(\tilde{t}|\tilde{\mathbf{x}})$. With equal class priors, this is equivalent to maximizing $p(\tilde{\mathbf{x}}|\tilde{t})$ and thus

$$\tilde{t} = \begin{cases} +1 & \text{iff } \dfrac{1}{N_{+1}} \displaystyle\sum_{i:t_i=+1} k(\tilde{\mathbf{x}}, \mathbf{x}_i) \geqslant \frac{1}{N_{-1}} \sum_{j:t_j=-1} k(\tilde{\mathbf{x}}, \mathbf{x}_j) \\ -1 & \text{otherwise.} \end{cases}$$

Here we have dropped the factor $1/Z_k$ since it only acts as a common scaling factor. Using the encoding scheme for the label, this classification rule can be written in the more compact form

$$\tilde{t} = \text{sign} \left(\sum_{n=1}^{N} \frac{t_n}{N_{t_n}} k(\tilde{\mathbf{x}}, \mathbf{x}_n) \right).$$

Now we take $k(\mathbf{x}, \mathbf{x}_n) = \mathbf{x}^{\mathrm{T}} \mathbf{x}_n$, which results in the kernel density

$$p(\mathbf{x}|t = +1) = \frac{1}{N_{+1}} \sum_{n:t_n=+1} \mathbf{x}^{\mathrm{T}} \mathbf{x}_n = \mathbf{x}^{\mathrm{T}} \bar{\mathbf{x}}^{+}.$$

Here, the sum in the middle experssion runs over all vectors $\mathbf{x}_n$ for which $t_n = +1$ and $\bar{\mathbf{x}}^{+}$ denotes the mean of these vectors, with the corresponding definition for the negative class. Note that this density is improper, since it cannot be normalized.

However, we can still compare likelihoods under this density, resulting in the classification rule

$$\tilde{t} = \begin{cases} +1 & \text{if } \tilde{\mathbf{x}}^{\mathrm{T}}\bar{\mathbf{x}}^{+} \geqslant \tilde{\mathbf{x}}^{\mathrm{T}}\bar{\mathbf{x}}^{-}, \\ -1 & \text{otherwise.} \end{cases}$$

The same argument would of course also apply in the feature space $\phi(\mathbf{x})$.

7.2 Consider multiplying both sides of (7.5) by $\gamma > 0$. Accordingly, we would then replace all occurences of $\mathbf{w}$ and b in (7.3) with $\gamma\mathbf{w}$ and γb, respectively. However, as discussed in the text following (7.3), its solution w.r.t. $\mathbf{w}$ and b is invariant to a common scaling factor and hence would remain unchanged.

7.3 Given a data set of two data points, $\mathbf{x}_1 \in \mathcal{C}_+$ ($t_1 = +1$) and $\mathbf{x}_2 \in \mathcal{C}_-$ ($t_2 = -1$), the maximum margin hyperplane is determined by solving (7.6) subject to the constraints

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 + b = +1 \tag{216}$$
$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 + b = -1. \tag{217}$$

We do this by introducing Lagrange multipliers λ and η, and solving

$$\arg\min_{\mathbf{w},b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + \lambda\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 + b - 1\right) + \eta\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 + b + 1\right) \right\}.$$

Taking the derivative of this w.r.t. $\mathbf{w}$ and b and setting the results to zero, we obtain

$$0 = \mathbf{w} + \lambda\mathbf{x}_1 + \eta\mathbf{x}_2 \tag{218}$$
$$0 = \lambda + \eta. \tag{219}$$

Equation (219) immediately gives $\lambda = -\eta$, which together with (218) give

$$\mathbf{w} = \lambda\left(\mathbf{x}_1 - \mathbf{x}_2\right). \tag{220}$$

For b, we first rearrange and sum (216) and (217) to obtain

$$2b = -\mathbf{w}^{\mathrm{T}}\left(\mathbf{x}_1 + \mathbf{x}_2\right).$$

Using (220), we can rewrite this as

$$\begin{aligned} b &= -\frac{\lambda}{2}\left(\mathbf{x}_1 - \mathbf{x}_2\right)^{\mathrm{T}}\left(\mathbf{x}_1 + \mathbf{x}_2\right) \\ &= -\frac{\lambda}{2}\left(\mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1 - \mathbf{x}_2^{\mathrm{T}}\mathbf{x}_2\right). \end{aligned}$$

Note that the Lagrange multiplier λ remains undetermined, which reflects the inherent indeterminacy in the magnitude of $\mathbf{w}$ and b.

7.4 From Figure 4.1 and (7.4), we see that the value of the margin

$$\rho = \frac{1}{\|\mathbf{w}\|} \quad \text{and so} \quad \frac{1}{\rho^2} = \|\mathbf{w}\|^2.$$

From (7.16) we see that, for the maximum margin solution, the second term of (7.7) vanishes and so we have

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2.$$

Using this together with (7.8), the dual (7.10) can be written as

$$\frac{1}{2}\|\mathbf{w}\|^2 = \sum_n a_n - \frac{1}{2}\|\mathbf{w}\|^2,$$

from which the desired result follows.

7.5 These properties follow directly from the results obtained in the solution to the previous exercise, 7.4.

7.6 If $p(t = 1|y) = \sigma(y)$, then

$$p(t = -1|y) \;=\; 1 - p(t = 1|y) \;=\; 1 - \sigma(y) \;=\; \sigma(-y),$$

where we have used (4.60). Thus, given i.i.d. data $\mathcal{D} = \{(t_1, \mathbf{x}_n), \ldots, (t_N, \mathbf{x}_N)\}$, we can write the corresponding likelihood as

$$p(\mathcal{D}) = \prod_{t_n=1} \sigma(y_n) \prod_{t_{n'}=-1} \sigma(-y_{n'}) = \prod_{n=1}^{N} \sigma(t_n y_n),$$

where $y_n = y(\mathbf{x}_n)$, as given by (7.1). Taking the negative logarithm of this, we get

$$\begin{aligned} -\ln p(\mathcal{D}) &= -\ln \prod_{n=1}^{N} \sigma(t_n y_n) \\ &= \sum_{n=1}^{N} \ln \sigma(t_n y_n) \\ &= \sum_{n=1}^{N} \ln(1 + \exp(-t_n y_n)), \end{aligned}$$

where we have used (4.59). Combining this with the regularization term $\lambda\|\mathbf{w}\|^2$, we obtain (7.47).

7.7 We start by rewriting (7.56) as

$$\begin{aligned} L &= \sum_{n=1}^{N} C\xi_n + \sum_{n=1}^{N} C\widehat{\xi}_n + \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{n=1}^{N}(\mu_n\xi_n + \widehat{\mu}_n\widehat{\xi}_n) \\ &\quad - \sum_{n=1}^{N} a_n(\epsilon + \xi_n + \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) + b - t_n) \\ &\quad - \sum_{n=1}^{N} \widehat{a}_n(\epsilon + \widehat{\xi}_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) - b + t_n), \end{aligned}$$

where we have used (7.1). We now use (7.1), (7.57), (7.59) and (7.60) to rewrite this as

$$\begin{aligned} L &= \sum_{n=1}^{N}(a_n + \mu_n)\xi_n + \sum_{n=1}^{N}(\widehat{a}_n + \widehat{\mu}_n)\widehat{\xi}_n \\ &\quad + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_m) - \sum_{n=1}^{N}(\mu_n\xi_n + \widehat{\mu}_n\widehat{\xi}_n) \\ &\quad - \sum_{n=1}^{N}(a_n\xi_n + \widehat{a}_n\widehat{\xi}_n) - \epsilon\sum_{n=1}^{N}(a_n + \widehat{a}_n) + \sum_{n=1}^{N}(a_n - \widehat{a}_n)t_n \\ &\quad - \sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_m) - b\sum_{n=1}^{N}(a_n - \widehat{a}_n). \end{aligned}$$

If we now eliminate terms that cancel out and use (7.58) to eliminate the last term, what we are left with equals the r.h.s. of (7.61).

7.8 This follows from (7.67) and (7.68), which in turn follow from the KKT conditions, (E.9)–(E.11), for μ_n, ξ_n, $\widehat{\mu}_n$ and $\widehat{\xi}_n$, and the results obtained in (7.59) and (7.60). For example, for μ_n and ξ_n, the KKT conditions are

$$\begin{aligned} \xi_n &\geqslant 0 \\ \mu_n &\geqslant 0 \\ \mu_n\xi_n &= 0 \end{aligned} \tag{221}$$

and from (7.59) we have that

$$\mu_n = C - a_n. \tag{222}$$

Combining (221) and (222), we get (7.67); similar reasoning for $\widehat{\mu}_n$ and $\widehat{\xi}_n$ lead to (7.68).

7.9 From (7.76), (7.79) and (7.80), we make the substitutions

$$\mathbf{x} \Rightarrow \mathbf{w} \quad \boldsymbol{\mu} \Rightarrow \mathbf{0} \quad \boldsymbol{\Lambda} \Rightarrow \mathrm{diag}(\boldsymbol{\alpha})$$

$$\mathbf{y} \Rightarrow \mathbf{t} \quad \mathbf{A} \Rightarrow \mathbf{\Phi} \quad \mathbf{b} \Rightarrow \mathbf{0} \quad \mathbf{L} \Rightarrow \beta\mathbf{I},$$

in (2.113) and (2.114), upon which the desired result follows from (2.116) and (2.117).

7.10 We first note that this result is given immediately from (2.113)–(2.115), but the task set in the exercise was to practice the technique of completing the square. In this solution and that of Exercise 7.12, we broadly follow the presentation in Section 3.5.1. Using (7.79) and (7.80), we can write (7.84) in a form similar to (3.78)

$$p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \frac{1}{(2\pi)^{N/2}} \prod_{i=1}^{M} \alpha_i \int \exp\{-E(\mathbf{w})\}\, \mathrm{d}\mathbf{w} \tag{223}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 + \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}$$

and $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$.

Completing the square over $\mathbf{w}$, we get

$$E(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{m})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{w} - \mathbf{m}) + E(\mathbf{t}) \tag{224}$$

where $\mathbf{m}$ and $\mathbf{\Sigma}$ are given by (7.82) and (7.83), respectively, and

$$E(\mathbf{t}) = \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{m}\right). \tag{225}$$

Using (224), we can evaluate the integral in (223) to obtain

$$\int \exp\{-E(\mathbf{w})\}\, \mathrm{d}\mathbf{w} = \exp\{-E(\mathbf{t})\}\,(2\pi)^{M/2}|\mathbf{\Sigma}|^{1/2}. \tag{226}$$

Considering this as a function of $\mathbf{t}$ we see from (7.83), that we only need to deal with the factor $\exp\{-E(\mathbf{t})\}$. Using (7.82), (7.83), (C.7) and (7.86), we can re-write (225) as follows

$$\begin{aligned}
E(\mathbf{t}) &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{m}\right) \\
&= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - \beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\beta\right) \\
&= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta\mathbf{I} - \beta\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^{\mathrm{T}}\beta\right)\mathbf{t} \\
&= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta\mathbf{I} - \beta\mathbf{\Phi}(\mathbf{A} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\beta\right)\mathbf{t} \\
&= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta^{-1}\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^{\mathrm{T}}\right)^{-1}\mathbf{t} \\
&= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t}.
\end{aligned}$$

This gives us the last term on the r.h.s. of (7.85); the two preceding terms are given implicitly, as they form the normalization constant for the posterior Gaussian distribution $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta)$.

7.11 If we make the same substitutions as in Exercise 7.9, the desired result follows from (2.115).

7.12 Using the results (223)–(226) from Solution 7.10, we can write (7.85) in the form of (3.86):

$$\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{N}{2}\ln\beta + \frac{1}{2}\sum_{i}^{N}\ln\alpha_i - E(\mathbf{t}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\ln(2\pi). \quad (227)$$

By making use of (225) and (7.83) together with (C.22), we can take the derivatives of this w.r.t α_i, yielding

$$\frac{\partial}{\partial\alpha_i}\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2}m_i^2. \quad (228)$$

Setting this to zero and re-arranging, we obtain

$$\alpha_i = \frac{1 - \alpha_i\Sigma_{ii}}{m_i^2} = \frac{\gamma_i}{m_i^2},$$

where we have used (7.89). Similarly, for β we see that

$$\frac{\partial}{\partial\beta}\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{1}{2}\left(\frac{N}{\beta} - \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 - \mathrm{Tr}\left[\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right]\right). \quad (229)$$

Using (7.83), we can rewrite the argument of the trace operator as

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} &= \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= \boldsymbol{\Sigma}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= (\mathbf{A} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= (\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma})\beta^{-1}. \quad (230)
\end{aligned}$$

Here the first factor on the r.h.s. of the last line equals (7.89) written in matrix form. We can use this to set (229) equal to zero and then re-arrange to obtain (7.88).

7.13 We start by introducing prior distributions over $\boldsymbol{\alpha}$ and β,

$$\begin{aligned}
p(\alpha_i) &= \mathrm{Gam}\left(\alpha_i|a_{\alpha 0}, b_{\beta 0}\right), i = 1, \ldots, N, \\
p(\beta) &= \mathrm{Gam}\left(\beta|a_{\beta 0}, b_{\beta 0}\right).
\end{aligned}$$

Note that we use an independent, common prior for all α_i. We can then combine this with (7.84) to obtain

$$p(\boldsymbol{\alpha}, \beta, \mathbf{t}|\mathbf{X}) = p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta)p(\boldsymbol{\alpha})p(\beta).$$

Rather than maximizing the r.h.s. directly, we first take the logarithm, which enables us to use results from Solution 7.12. Using (227) and (B.26), we get

$$\begin{aligned}\ln p(\boldsymbol{\alpha}, \beta, \mathbf{t}|\mathbf{X}) = &\frac{N}{2}\ln\beta + \frac{1}{2}\sum_{i}^{N}\ln\alpha_i - E(\mathbf{t}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\ln(2\pi) \\ &-N\ln\Gamma(a_{\alpha 0})^{-1} + Na_{\alpha 0}\ln b_{\alpha 0} + \sum_{i=1}^{N}\left((a_{\alpha 0}-1)\ln\alpha_i - b_{\alpha 0}\alpha_i\right) \\ &-\ln\Gamma(a_{\beta 0})^{-1} + a_{\beta 0}\ln b_{\beta 0} + (a_{\beta 0}-1)\ln\beta - b_{\beta 0}\beta.\end{aligned}$$

Using (228), we obtain the derivative of this w.r.t. α_i as

$$\frac{\partial}{\partial\alpha_i}\ln p(\boldsymbol{\alpha}, \beta, \mathbf{t}|\mathbf{X}) = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2}m_i^2 + \frac{a_{\alpha 0}-1}{\alpha_i} - b_{\alpha 0}.$$

Setting this to zero and rearranging (cf. Solution 7.12) we obtain

$$\alpha_i^{\text{new}} = \frac{\gamma_i + 2a_{\alpha 0} - 2}{m_i^2 - 2b_{\alpha 0}},$$

where we have used (7.89).

For β, we can use (229) together with (B.26) to get

$$\frac{\partial}{\partial\beta}\ln p(\boldsymbol{\alpha}, \beta, \mathbf{t}|\mathbf{X}) = \frac{1}{2}\left(\frac{N}{\beta} - \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 - \text{Tr}\left[\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Phi}\right]\right) + \frac{a_{\beta 0}-1}{\beta} - b_{\beta 0}.$$

Setting this equal to zero and using (7.89) and (230), we get

$$\frac{1}{\beta^{\text{new}}} = \frac{\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 + 2b_{\beta 0}}{a_{\beta 0} + 2 + N - \sum_i \gamma_i}.$$

7.14 If we make the following substitions from (7.81) into (2.113),

$$\mathbf{x} \Rightarrow \mathbf{w} \quad \boldsymbol{\mu} \Rightarrow \mathbf{m} \quad \boldsymbol{\Lambda}^{-1} \Rightarrow \boldsymbol{\Sigma},$$

and from (7.76) and (7.77) into (2.114)

$$\mathbf{y} \Rightarrow t \quad \mathbf{A} \Rightarrow \boldsymbol{\phi}(\mathbf{x})^{\text{T}} \quad \mathbf{b} \Rightarrow \mathbf{0} \quad \mathbf{L} \Rightarrow \beta^{\star}\mathbf{I},$$

(7.90) and (7.91) can be read off directly from (2.115).

7.15 Using (7.94), (7.95) and (7.97)–(7.99), we can rewrite (7.85) as follows

$$\begin{aligned}
\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) &= -\frac{1}{2}\Big\{ N\ln(2\pi) + \ln|\mathbf{C}_{-i}||1+\alpha_i^{-1}\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| \\
&\quad + \mathbf{t}^{\mathrm{T}}\left(\mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}\right)\mathbf{t}\Big\} \\
&= -\frac{1}{2}\left\{N\ln(2\pi) + \ln|\mathbf{C}_{-i}| + \mathbf{t}^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\mathbf{t}\right\} \\
&\quad + \frac{1}{2}\left[-\ln|1+\alpha_i^{-1}\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| + \mathbf{t}^{\mathrm{T}}\frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}\mathbf{t}\right] \\
&= L(\alpha_{-i}) + \frac{1}{2}\left[\ln\alpha_i - \ln(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right] \\
&= L(\alpha_{-i}) + \lambda(\alpha_i)
\end{aligned}$$

7.16 If we differentiate (7.97) twice w.r.t. α_i, we get

$$\frac{\mathrm{d}^2\lambda}{\mathrm{d}\alpha_i^2} = -\frac{1}{2}\left(\frac{1}{\alpha_i^2} + \frac{1}{(\alpha_i + s_i)^2}\right).$$

This second derivative must be negative and thus the solution given by (7.101) corresponds to a maximum.

7.17 Using (7.83), (7.86) and (C.7), we have

$$\mathbf{C}^{-1} = \beta\mathbf{I} - \beta^2\boldsymbol{\Phi}\left(\mathbf{A} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}} = \beta\mathbf{I} - \beta^2\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}.$$

Substituting this into (7.102) and (7.103), we immediately obtain (7.106) and (7.107), respectively.

7.18 As the RVM can be regarded as a regularized logistic regression model, we can follow the sequence of steps used to derive (4.91) in Exercise 4.13 to derive the first term of the r.h.s. of (7.110), whereas the second term follows from standard matrix derivatives (see Appendix C). Note however, that in Exercise 4.13 we are dealing with the *negative* log-likelihood.

To derive (7.111), we make use of (161) and (162) from Exercise 4.13. If we write the first term of the r.h.s. of (7.110) in component form we get

$$\begin{aligned}
\frac{\partial}{\partial w_j}\sum_{n=1}^{N}(t_n - y_n)\phi_{ni} &= -\sum_{n=1}^{N}\frac{\partial y_n}{\partial a_n}\frac{\partial a_n}{\partial w_j}\phi_{ni} \\
&= -\sum_{n=1}^{N} y_n(1-y_n)\phi_{nj}\phi_{ni},
\end{aligned}$$

which, written in matrix form, equals the first term inside the parenthesis on the r.h.s. of (7.111). The second term again follows from standard matrix derivatives.

7.19 **NOTE**: In the 1st printing of PRML, on line 1 of the text of this exercise, "approximate log marginal" should be "approximate marginal".

We start by taking the logarithm of (7.114), which, omitting terms that do not depend on $\boldsymbol{\alpha}$, leaves us with

$$\ln p(\mathbf{w}^\star|\boldsymbol{\alpha}) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| = -\frac{1}{2}\left(\ln|\boldsymbol{\Sigma}^{-1}| + \sum_i (w_i^\star)^2\,\alpha_i - \ln\alpha_i\right),$$

where we have used (7.80). Making use of (7.113) and (C.22), we can differentiate this to obtain (7.115), from which we get (7.116) by using $\gamma_i = 1 - \alpha_i\Sigma_{ii}$.

Chapter 8 Graphical Models

8.1 We want to show that, for (8.5),

$$\sum_{x_1}\cdots\sum_{x_K} p(\mathbf{x}) = \sum_{x_1}\cdots\sum_{x_K}\prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) = 1.$$

We assume that the nodes in the graph has been numbered such that x_1 is the root node and no arrows lead from a higher numbered node to a lower numbered node. We can then marginalize over the nodes in reverse order, starting with x_K

$$\begin{aligned}\sum_{x_1}\cdots\sum_{x_K} p(\mathbf{x}) &= \sum_{x_1}\cdots\sum_{x_K} p(x_K|\mathrm{pa}_K)\prod_{k=1}^{K-1} p(x_k|\mathrm{pa}_k)\\ &= \sum_{x_1}\cdots\sum_{x_{K-1}}\prod_{k=1}^{K-1} p(x_k|\mathrm{pa}_k),\end{aligned}$$

since each of the conditional distributions is assumed to be correctly normalized and none of the other variables depend on x_K. Repeating this process $K-2$ times we are left with

$$\sum_{x_1} p(x_1|\emptyset) = 1.$$

8.2 Consider a directed graph in which the nodes of the graph are numbered such that are no edges going from a node to a lower numbered node. If there exists a directed cycle in the graph then the subset of nodes belonging to this directed cycle must also satisfy the same numbering property. If we traverse the cycle in the direction of the edges the node numbers cannot be monotonically increasing since we must end up back at the starting node. It follows that the cycle cannot be a directed cycle.

Table 1 Comparison of the distribution $p(a, b)$ with the product of marginals $p(a)p(b)$ showing that these are not equal for the given joint distribution $p(a, b, c)$.

a	b	$p(a, b)$
0	0	336.000
0	1	264.000
1	0	256.000
1	1	144.000

a	b	$p(a)p(b)$
0	0	355200.000
0	1	244800.000
1	0	236800.000
1	1	163200.000

8.3 The distribution $p(a, b)$ is found by summing the complete joint distribution $p(a, b, c)$ over the states of c so that

$$p(a, b) = \sum_{c\in\{0,1\}} p(a, b, c)$$

and similarly the marginal distributions $p(a)$ and $p(b)$ are given by

$$p(a) = \sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} p(a, b, c) \quad \text{and} \quad p(b) = \sum_{a\in\{0,1\}} \sum_{c\in\{0,1\}} p(a, b, c). \tag{231}$$

Table 1 shows the joint distribution $p(a, b)$ as well as the product of marginals $p(a)p(b)$, demonstrating that these are not equal for the specified distribution.

The conditional distribution $p(a, b|c)$ is obtained by conditioning on the value of c and normalizing

$$p(a, b|c) = \frac{p(a, b, c)}{\sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a, b, c)}.$$

Similarly for the conditionals $p(a|c)$ and $p(b|c)$ we have

$$p(a|c) = \frac{\sum_{b\in\{0,1\}} p(a, b, c)}{\sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a, b, c)}$$

and

$$p(b|c) = \frac{\sum_{a\in\{0,1\}} p(a, b, c)}{\sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a, b, c)}. \tag{232}$$

Table 2 compares the conditional distribution $p(a, b|c)$ with the product of marginals $p(a|c)p(b|c)$, showing that these are equal for the given joint distribution $p(a, b, c)$ for both $c = 0$ and $c = 1$.

8.4 In the previous exercise we have already computed $p(a)$ in (231) and $p(b|c)$ in (232). There remains to compute $p(c|a)$ which is done using

$$p(c|a) = \frac{\sum_{b\in\{0,1\}} p(a, b, c)}{\sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} p(a, b, c)}.$$

Table 2 Comparison of the conditional distribution $p(a, b|c)$ with the product of marginals $p(a|c)p(b|c)$ showing that these are equal for the given distribution.

a	b	c	$p(a, b\|c)$
0	0	0	0.400
0	1	0	0.100
1	0	0	0.400
1	1	0	0.100
0	0	1	0.277
0	1	1	0.415
1	0	1	0.123
1	1	1	0.185

a	b	c	$p(a\|c)p(b\|c)$
0	0	0	0.400
0	1	0	0.100
1	0	0	0.400
1	1	0	0.100
0	0	1	0.277
0	1	1	0.415
1	0	1	0.123
1	1	1	0.185

The required distributions are given in Table 3.

Table 3 Tables of $p(a)$, $p(c|a)$ and $p(b|c)$ evaluated by marginalizing and conditioning the joint distribution of Table 8.2.

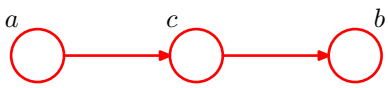
a	$p(a)$
0	600.000
1	400.000

c	a	$p(c\|a)$
0	0	0.400
1	0	0.600
0	1	0.600
1	1	0.400

b	c	$p(b\|c)$
0	0	0.800
1	0	0.200
0	1	0.400
1	1	0.600

Multiplying the three distributions together we recover the joint distribution $p(a, b, c)$ given in Table 8.2, thereby allowing us to verify the validity of the decomposition $p(a, b, c) = p(a)p(c|a)p(b|c)$ for this particular joint distribution. We can express this decomposition using the graph shown in Figure 4.

Figure 4 Directed graph representing the joint distribution given in Table 8.2.



8.5 **NOTE**: In PRML, Equation (7.79) contains a typographical error: $p(t_n|\mathbf{x}_n, \mathbf{w}, \beta^{-1})$ should be $p(t_n|\mathbf{x}_n, \mathbf{w}, \beta)$. This correction is provided for completeness only; it does not affect this solution.

The solution is given in Figure 5.

8.6 **NOTE**: In PRML, the text of the exercise should be slightly altered; please consult the PRML errata.

In order to interpret (8.104) suppose initially that $\mu_0 = 0$ and that $\mu_i = 1 - \epsilon$ where $\epsilon \ll 1$ for $i = 1, \ldots, K$. We see that, if all of the $x_i = 0$ then $p(y = 1|x_1, \ldots, x_K) = 0$ while if L of the $x_i = 1$ then $p(y = 1|x_1, \ldots, x_K) = 1 - \epsilon^L$ which is close to 1. For $\epsilon \to 0$ this represents the logical OR function in which $y = 1$ if one or more of the $x_i = 1$, and $y = 0$ otherwise. More generally, if just one of the $x_i = 1$ with all remaining $x_{j \neq i} = 0$ then $p(y = 1|x_1, \ldots, x_K) = \mu_i$ and so we can interpret μ_i as the probability of $y = 1$ given that only this one $x_i = 1$. We can similarly interpret μ_0 as the probability of $y = 1$ when all of the $x_i = 0$. An example of the application of this model would be in medical diagnosis in which y

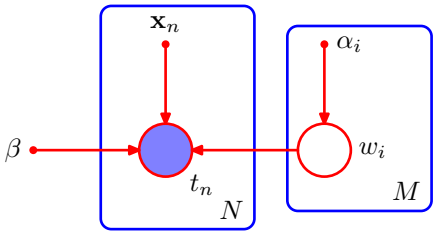


Figure 5 The graphical representation of the relevance vector machine (RVM); Solution 8.5.

represents the presence or absence of a symptom, and each of the x_i represents the presence or absence of some disease. For the i^{th} disease there is a probability μ_i that it will give rise to the symptom. There is also a background probability μ_0 that the symptom will be observed even in the absence of disease. In practice we might observe that the symptom is indeed present (so that $y = 1$) and we wish to infer the posterior probability for each disease. We can do this using Bayes' theorem once we have defined prior probabilities $p(x_i)$ for the diseases.

8.7 Starting with $\boldsymbol{\mu}$, (8.11) and (8.15) directly gives

$$\mu_1 = \sum_{j \in \emptyset} w_{1j}\mathbb{E}[x_j] + b_1 = b_1,$$

$$\mu_2 = \sum_{j \in \{x_1\}} w_{2j}\mathbb{E}[x_j] + b_2 = w_{21}b_1 + b_2$$

and

$$\mu_3 = \sum_{j \in \{x_2\}} w_{3j}\mathbb{E}[x_j] + b_3 = w_{32}(w_{21}b_1 + b_2) + b_3.$$

Similarly for $\boldsymbol{\Sigma}$, using (8.11) and (8.16), we get

$$\text{cov}[x_1, x_1] = \sum_{k \in \emptyset} w_{1j}\text{cov}[x_1, x_k] + I_{11}v_1 = v_1,$$

$$\text{cov}[x_1, x_2] = \sum_{k \in \{x_1\}} w_{2j}\text{cov}[x_1, x_k] + I_{12}v_2 = w_{21}v_1,$$

$$\text{cov}[x_1, x_3] = \sum_{k \in \{x_2\}} w_{3j}\text{cov}[x_1, x_k] + I_{13}v_3 = w_{32}w_{21}v_1,$$

$$\text{cov}[x_2, x_2] = \sum_{k \in \{x_1\}} w_{2j}\text{cov}[x_2, x_k] + I_{22}v_2 = w_{21}^2 v_1 + v_2,$$

$$\text{cov}[x_2, x_3] = \sum_{k \in \{x_2\}} w_{3j}\text{cov}[x_2, x_k] + I_{23}v_3 = w_{32}(w_{21}^2 v_1 + v_2)$$

and

$$\text{cov}[x_3, x_3] = \sum_{k \in \{x_2\}} w_{3j}\text{cov}[x_3, x_k] + I_{33}v_3 = w_{32}^2(w_{21}^2 v_1 + v_2) + v_3,$$

where the symmetry of $\boldsymbol{\Sigma}$ gives the below diagonal elements.

8.8 $a \perp\!\!\!\perp b, c \mid d$ can be written as

$$p(a, b, c|d) = p(a|d)p(b, c|d).$$

Summing (or integrating) both sides with respect to c, we obtain

$$p(a, b|d) = p(a|d)p(b|d) \qquad \text{or} \qquad a \perp\!\!\!\perp b \mid d,$$

as desired.

8.9 Consider Figure 8.26. In order to apply the d-separation criterion we need to consider all possible paths from the central node $\mathbf{x}_i$ to all possible nodes external to the Markov blanket. There are three possible categories of such paths. First, consider paths via the parent nodes. Since the link from the parent node to the node $\mathbf{x}_i$ has its tail connected to the parent node, it follows that for any such path the parent node must be either tail-to-tail or head-to-tail with respect to the path. Thus the observation of the parent node will block any such path. Second consider paths via one of the child nodes of node $\mathbf{x}_i$ which do not pass directly through any of the co-parents. By definition such paths must pass to a child of the child node and hence will be head-to-tail with respect to the child node and so will be blocked. The third and final category of path passes via a child node of $\mathbf{x}_i$ and then a co-parent node. This path will be head-to-head with respect to the observed child node and hence will not be blocked by the observed child node. However, this path will either tail-to-tail or head-to-tail with respect to the co-parent node and hence observation of the co-parent will block this path. We therefore see that all possible paths leaving node $\mathbf{x}_i$ will be blocked and so the distribution of $\mathbf{x}_i$, conditioned on the variables in the Markov blanket, will be independent of all of the remaining variables in the graph.

8.10 From Figure 8.54, we see that

$$p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c).$$

Following the examples in Section 8.2.1, we see that

$$\begin{aligned} p(a, b) &= \sum_c \sum_d p(a, b, c, d) \\ &= p(a)p(b) \sum_c p(c|a, b) \sum_d p(d|c) \\ &= p(a)p(b). \end{aligned}$$

Similarly,

$$\begin{aligned} p(a,b|d) &= \frac{\sum_c p(a,b,c,d)}{\sum_a \sum_b \sum_c p(a,b,c,d)} \\ &= \frac{p(d|a,b)p(a)p(b)}{p(d)} \\ &\neq p(a|d)p(b|d) \end{aligned}$$

in general. Note that this result could also be obtained directly from the graph in Figure 8.54 by using d-separation, discussed in Section 8.2.2.

8.11 The described situation correspond to the graph shown in Figure 8.54 with $a = B$, $b = F$, $c = G$ and $d = D$ (cf. Figure 8.21). To evaulate the probability that the tank is empty given the driver's report that the gauge reads zero, we use Bayes' theorem

$$p(F = 0|D = 0) = \frac{p(D = 0|F = 0)p(F = 0)}{p(D = 0)}.$$

To evaluate $p(D = 0|F = 0)$, we marginalize over B and G,

$$p(D = 0|F = 0) = \sum_{B,G} p(D = 0|G)p(G|B, F = 0)p(B) = 0.748 \tag{233}$$

and to evaluate $p(D = 0)$, we marginalize also over F,

$$p(D = 0) = \sum_{B,G,F} p(D = 0|G)p(G|B, F)p(B)p(F) = 0.352. \tag{234}$$

Combining these results with $p(F = 0)$, we get

$$p(F = 0|D = 0) = 0.213.$$

Note that this is slightly lower than the probability obtained in (8.32), reflecting the fact that the driver is not completely reliable.

If we now also observe $B = 0$, we longer marginalize over B in (233) and (234), but instead keep it fixed at its observed value, yielding

$$p(F = 0|D = 0, B = 0) = 0.110$$

which is again lower than what we obtained with a direct observation of the fuel gauge in (8.33). More importantly, in both cases the value is lower than before we observed $B = 0$, since this observation provides an alternative explanation why the gauge should read zero; see also discussion following (8.33).

8.12 In an undirected graph of M nodes there could potentially be a link between each pair of nodes. The number of distinct graphs is then 2 raised to the power of the number of potential links. To evaluate the number of distinct links, note that there

are M nodes each of which could have a link to any of the other $M-1$ nodes, making a total of $M(M-1)$ links. However, each link is counted twice since, in an undirected graph, a link from node a to node b is equivalent to a link from node b to node a. The number of distinct potential links is therefore $M(M-1)/2$ and so the number of distinct graphs is $2^{M(M-1)/2}$. The set of 8 possible graphs over three nodes is shown in Figure 6.

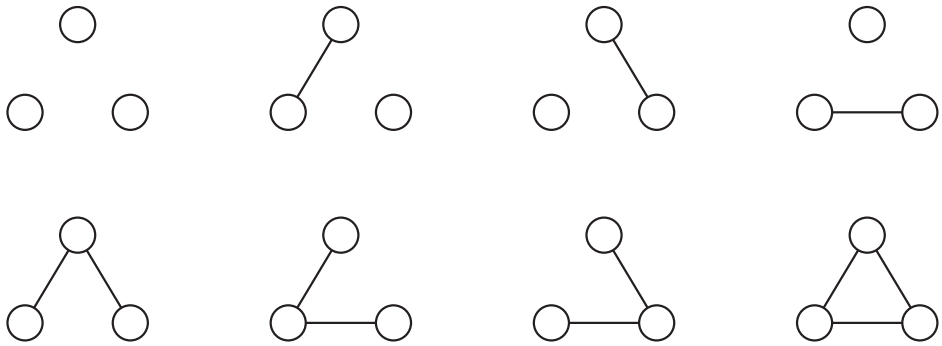


Figure 6 The set of 8 distinct undirected graphs which can be constructed over $M = 3$ nodes.

8.13 The change in energy is

$$E(x_j = +1) - E(x_j = -1) = 2h - 2\beta \sum_{i \in \mathrm{ne}(j)} x_i - 2\eta y_j$$

where $\mathrm{ne}(j)$ denotes the nodes which are neighbours of x_j.

8.14 The most probable configuration corresponds to the configuration with the lowest energy. Since η is a positive constant (and $h = \beta = 0$) and $x_i, y_i \in \{-1, +1\}$, this will be obtained when $x_i = y_i$ for all $i = 1, \ldots, D$.

8.15 The marginal distribution $p(x_{n-1}, x_n)$ is obtained by marginalizing the joint distribution $p(\mathbf{x})$ over all variables except x_{n-1} and x_n,

$$p(x_{n-1}, x_n) = \sum_{x_1} \cdots \sum_{x_{n-2}} \sum_{x_{n+1}} \cdots \sum_{x_N} p(\mathbf{x}).$$

This is analogous to the marginal distribution for a single variable, given by (8.50).

Following the same steps as in the single variable case described in Section 8.4.1,

we arrive at a modified form of (8.52),

$$p(x_n) = \frac{1}{Z} \underbrace{\left[\sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \cdots \left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right] \cdots\right] \psi_{n-1,n}(x_{n-1}, x_n)}_{\mu_\alpha(x_{n-1})}$$
$$\underbrace{\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \cdots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots\right]}_{\mu_\beta(x_n)},$$

from which (8.58) immediately follows.

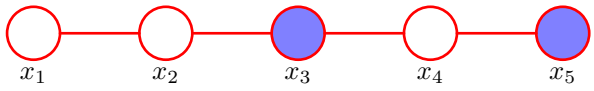
8.16 Observing $\mathbf{x}_N = \widehat{\mathbf{x}}_N$ will only change the initial expression (message) for the β-recursion, which now becomes

$$\mu_\beta(\mathbf{x}_{N-1}) = \psi_{N-1,N}(\mathbf{x}_{N-1}, \widehat{\mathbf{x}}_N).$$

Note that there is no summation over $\mathbf{x}_N$. $p(\mathbf{x}_n)$ can then be evaluated using (8.54)–(8.57) for all $n = 1, \ldots, N-1$.

8.17 With $N = 5$ and x_3 and x_5 observed, the graph from Figure 8.38 will look like in Figure 7. This graph is undirected, but from Figure 8.32 we see that the equivalent

Figure 7 The graph discussed in Solution 8.17.



directed graph can be obtained by simply directing all the edges from left to right. (**NOTE**: In PRML, the labels of the two rightmost nodes in Figure 8.32b should be interchanged to be the same as in Figure 8.32a.) In this directed graph, the edges on the path from x_2 to x_5 meet head-to-tail at x_3 and since x_3 is observed, by d-separation $x_2 \perp\!\!\!\perp x_5 | x_3$; note that we would have obtained the same result if we had chosen to direct the arrows from right to left. Alternatively, we could have obtained this result using graph separation in undirected graphs, illustrated in Figure 8.27.

From (8.54), we have

$$p(x_2) = \frac{1}{Z} \mu_\alpha(x_2) \mu_\beta(x_2). \tag{235}$$

$\mu_\alpha(x_2)$ is given by (8.56), while for $\mu_\beta(x_2)$, (8.57) gives

$$\begin{aligned} \mu_\beta(x_2) &= \sum_{x_3} \psi_{2,3}(x_2, x_3) \mu_\beta(x_3) \\ &= \psi_{2,3}(x_2, \widehat{x}_3) \mu_\beta(\widehat{x}_3) \end{aligned}$$

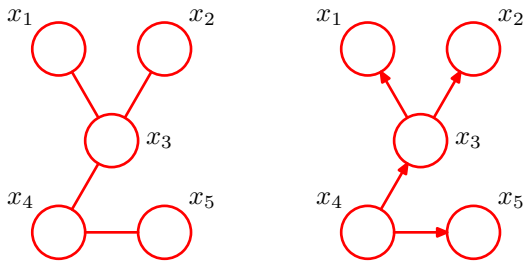


Figure 8 The graph on the left is an undirected tree. If we pick x_4 to be the root node and direct all the edges in the graph to point from the root to the leaf nodes (x_1, x_2 and x_5), we obtain the directed tree shown on the right.

since x_3 is observed and we denote the observed value $\widehat{x}_3$. Thus, any influence that x_5 might have on $\mu_\beta(\widehat{x}_3)$ will be in terms of a scaling factor that is indepedent of x_2 and which will be absorbed into the normalization constant Z in (235) and so

$$p(x_2|x_3, x_5) = p(x_2|x_3).$$

8.18 The joint probability distribution over the variables in a general directed graphical model is given by (8.5). In the particular case of a tree, each node has a single parent, so pa_k will be a singleton for each node, k, except for the root node for which it will empty. Thus, the joint probability distribution for a tree will be similar to the joint probability distribution over a chain, (8.44), with the difference that the same variable may occur to the right of the conditioning bar in several conditional probability distributions, rather than just one (in other words, although each node can only have one parent, it can have several children). Hence, the argument in Section 8.3.4, by which (8.44) is re-written as (8.45), can also be applied to probability distributions over trees. The result is a Markov random field model where each potential function corresponds to one conditional probability distribution in the directed tree. The prior for the root node, e.g. $p(x_1)$ in (8.44), can again be incorporated in one of the potential functions associated with the root node or, alternatively, can be incorporated as a single node potential.

This transformation can also be applied in the other direction. Given an undirected tree, we pick a node arbitrarily as the root. Since the graph is a tree, there is a unique path between every pair of nodes, so, starting at root and working outwards, we can direct all the edges in the graph to point from the root to the leaf nodes. An example is given in Figure 8. Since every edge in the tree correspond to a two-node potential function, by normalizing this appropriately, we obtain a conditional probability distribution for the child given the parent.

Since there is a unique path beween every pair of nodes in an undirected tree, once we have chosen the root node, the remainder of the resulting directed tree is given. Hence, from an undirected tree with N nodes, we can construct N different directed trees, one for each choice of root node.

8.19 If we convert the chain model discussed in Section 8.4.1 into a factor graph, each potential function in (8.49) will become a factor. Under this factor graph model, $p(x_n)$ is given by (8.63) as

$$p(x_n) = \mu_{f_{n-1,n}\to x_n}(x_n)\mu_{f_{n,n+1}\to x_n}(x_n) \tag{236}$$

where we have adopted the indexing of potential functions from (8.49) to index the factors. From (8.64)–(8.66), we see that

$$\mu_{f_{n-1,n}\to x_n}(x_n) = \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\mu_{x_{n-1}\to f_{n-1,n}}(x_{n-1}) \tag{237}$$

and

$$\mu_{f_{n,n+1}\to x_n}(x_n) = \sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1})\mu_{x_{n+1}\to f_{n,n+1}}(x_{n+1}). \tag{238}$$

From (8.69), we further see that

$$\mu_{x_{n-1}\to f_{n-1,n}}(x_{n-1}) = \mu_{f_{n-2,n-1}\to x_{n-1}}(x_{n-1})$$

and

$$\mu_{x_{n+1}\to f_{n,n+1}}(x_{n+1}) = \mu_{f_{n+1,n+2}\to x_{n+1}}(x_{n+1}).$$

Substituting these into (237) and (238), respectively, we get

$$\mu_{f_{n-1,n}\to x_n}(x_n) = \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\mu_{f_{n-2,n-1}\to x_{n-1}}(x_{n-1}) \tag{239}$$

and

$$\mu_{f_{n,n+1}\to x_n}(x_n) = \sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1})\mu_{f_{n+1,n+2}\to x_{n+1}}(x_{n+1}). \tag{240}$$

Since the messages are uniquely identified by the index of their arguments and whether the corresponding factor comes before or after the argument node in the chain, we can rename the messages as

$$\mu_{f_{n-2,n-1}\to x_{n-1}}(x_{n-1}) = \mu_\alpha(x_{n-1})$$

and

$$\mu_{f_{n+1,n+2}\to x_{n+1}}(x_{n+1}) = \mu_\beta(x_{n+1}).$$

Applying these name changes to both sides of (239) and (240), respectively, we recover (8.55) and (8.57), and from these and (236) we obtain (8.54); the normalization constant $1/Z$ can be easily computed by summing the (unnormalized) r.h.s. of (8.54). Note that the end nodes of the chain are variable nodes which send unit messages to their respective neigbouring factors (cf. (8.56)).

8.20 We do the induction over the size of the tree and we grow the tree one node at a time while, at the same time, we update the message passing schedule. Note that we can build up any tree this way.

For a single root node, the required condition holds trivially true, since there are no messages to be passed. We then assume that it holds for a tree with N nodes. In the induction step we add a new leaf node to such a tree. This new leaf node need not

to wait for any messages from other nodes in order to send its outgoing message and so it can be scheduled to send it first, before any other messages are sent. Its parent node will receive this message, whereafter the message propagation will follow the schedule for the original tree with N nodes, for which the condition is assumed to hold.

For the propagation of the outward messages from the root back to the leaves, we first follow the propagation schedule for the original tree with N nodes, for which the condition is assumed to hold. When this has completed, the parent of the new leaf node will be ready to send its outgoing message to the new leaf node, thereby completing the propagation for the tree with $N + 1$ nodes.

8.21 **NOTE**: In the 1st printing of PRML, this exercise contains a typographical error. On line 2, $f_x(\mathbf{x}_s)$ should be $f_s(\mathbf{x}_s)$.

To compute $p(\mathbf{x}_s)$, we marginalize $p(\mathbf{x})$ over all other variables, analogously to (8.61),

$$p(\mathbf{x}_s) = \sum_{\mathbf{x}\setminus\mathbf{x}_s} p(\mathbf{x}).$$

Using (8.59) and the defintion of $F_s(x, X_s)$ that followed (8.62), we can write this as

$$\begin{aligned} p(\mathbf{x}_s) &= \sum_{\mathbf{x}\setminus\mathbf{x}_s} f_s(\mathbf{x}_s) \prod_{i\in\mathrm{ne}(f_s)} \prod_{j\in\mathrm{ne}(x_i)\setminus f_s} F_j(x_i, X_{ij}) \\ &= f_s(\mathbf{x}_s) \prod_{i\in\mathrm{ne}(f_s)} \sum_{\mathbf{x}\setminus\mathbf{x}_s} \prod_{j\in\mathrm{ne}(x_i)\setminus f_s} F_j(x_i, X_{ij}) \\ &= f_s(\mathbf{x}_s) \prod_{i\in\mathrm{ne}(f_s)} \mu_{x_i\to f_s}(x_i), \end{aligned}$$

where in the last step, we used (8.67) and (8.68). Note that the marginalization over the different sub-trees rooted in the neighbours of f_s would only run over variables in the respective sub-trees.

8.22 Let X_a denote the set of variable nodes in the connected subgraph of interest and X_b the remaining variable nodes in the full graph. To compute the joint distribution over the variables in X_a, we need to marginalize $p(\mathbf{x})$ over X_b,

$$p(X_a) = \sum_{X_b} p(\mathbf{x}).$$

We can use the sum-product algorithm to perform this marginalization efficiently, in the same way that we used it to marginalize over all variables but x_n when computing $p(x_n)$. Following the same steps as in the single variable case (see Section 8.4.4),

we can write can write $p(X_a)$ in a form corresponding to (8.63),

$$\begin{aligned} p(X_a) &= \prod_{s_a} f_{s_a}(X_{s_a}) \prod_{s \in \mathrm{ne} X_a} \sum_{X_s} F_s(x_s, X_s) \\ &= \prod_{s_a} f_{s_a}(X_{s_a}) \prod_{s \in \mathrm{ne} X_a} \mu_{f_s \to x_s}(x_s). \end{aligned} \tag{241}$$

Here, s_a indexes factors that only depend on variables in X_a and so $X_{s_a} \subseteq X_a$ for all values of s_a; s indexes factors that connect X_a and X_b and hence also the corresponding nodes, $x_s \in X_a$. $X_s \subseteq X_b$ denotes the variable nodes connected to x_s via factor f_s. The messages $\mu_{f_s \to x_s}(x_s)$ can be computed using the sum-product algorithm, starting from the leaf nodes in, or connected to nodes in, X_b. Note that the density in (241) may require normalization, which will involve summing the r.h.s. of (241) over all possible combination of values for X_a.

8.23 This follows from the fact that the message that a node, x_i, will send to a factor f_s, consists of the product of all other messages received by x_i. From (8.63) and (8.69), we have

$$\begin{aligned} p(x_i) &= \prod_{s \in \mathrm{ne}(x_i)} \mu_{f_s \to x_i}(x_i) \\ &= \mu_{f_s \to x_i}(x_i) \prod_{t \in \mathrm{ne}(x_i) \setminus f_s} \mu_{f_t \to x_i}(x_i) \\ &= \mu_{f_s \to x_i}(x_i)\, \mu_{x_i \to f_s}(x_i). \end{aligned}$$

8.24 **NOTE**: In PRML, this exercise contains a typographical error. On the last line, $f(\mathbf{x}_s)$ should be $f_s(\mathbf{x}_s)$.

See Solution 8.21.

8.25 **NOTE**: In the 1st printing of PRML, equation (8.86) contains a typographical error. On the third line, the second summation should sum over x_3, not x_2. Furthermore, in equation (8.79), "$\mu_{x_2 \to f_b}$" (no argument) should be "$\mu_{x_2 \to f_b}(x_2)$".

Starting from (8.63), using (8.73), (8.77) and (8.81)–(8.83), we get

$$\begin{aligned}
\widetilde{p}(x_1) &= \mu_{f_a \to x_1}(x_1) \\
&= \sum_{x_2} f_a(x_1, x_2) \mu_{x_2 \to f_a}(x_2) \\
&= \sum_{x_2} f_a(x_1, x_2) \mu_{f_b \to x_2}(x_2) \mu_{f_c \to x_2}(x_2) \\
&= \sum_{x_2} f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_c(x_2, x_4) \\
&= \sum_{x_2} \sum_{x_3} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) \\
&= \sum_{x_2} \sum_{x_3} \sum_{x_4} \widetilde{p}(\mathbf{x}).
\end{aligned}$$

Similarly, starting from (8.63), using (8.73), (8.75) and (8.77)–(8.79), we get

$$\begin{aligned}
\widetilde{p}(x_3) &= \mu_{f_b \to x_3}(x_3) \\
&= \sum_{x_2} f_b(x_2, x_3) \mu_{x_2 \to f_b}(x_2) \\
&= \sum_{x_2} f_b(x_2, x_3) \mu_{f_a \to x_2}(x_2) \mu_{f_c \to x_2}(x_2) \\
&= \sum_{x_2} f_b(x_2, x_3) \sum_{x_1} f_a(x_1, x_2) \sum_{x_4} f_c(x_2, x_4) \\
&= \sum_{x_1} \sum_{x_2} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) \\
&= \sum_{x_1} \sum_{x_2} \sum_{x_4} \widetilde{p}(\mathbf{x}).
\end{aligned}$$

Finally, starting from (8.72), using (8.73), (8.74), (8.77), (8.81) and (8.82), we get

$$\begin{aligned}
\widetilde{p}(x_1, x_2) &= f_a(x_1, x_2) \mu_{x_1 \to f_a}(x_1) \mu_{x_2 \to f_a}(x_2) \\
&= f_a(x_1, x_2) \mu_{f_b \to x_2}(x_2) \mu_{f_c \to x_2}(x_2) \\
&= f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_b(x_2, x_4) \\
&= \sum_{x_3} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_b(x_2, x_4) \\
&= \sum_{x_3} \sum_{x_4} \widetilde{p}(\mathbf{x}).
\end{aligned}$$

8.26 We start by using the product and sum rules to write

$$p(x_a, x_b) = p(x_b|x_a)p(x_a) = \sum_{\mathbf{x}_{\setminus ab}} p(\mathbf{x}) \tag{242}$$

where $\mathbf{x}_{\setminus ab}$ denote the set of all all variables in the graph except x_a and x_b.

We can use the sum-product algorithm from Section 8.4.4 to first evaluate $p(x_a)$, by marginalizing over all other variables (including x_b). Next we successively fix x_a at all its allowed values and for each value, we use the sum-product algorithm to evaluate $p(x_b|x_a)$, by marginalizing over all variables except x_b and x_a, the latter of which will only appear in the formulae at its current, fixed value. Finally, we use (242) to evaluate the joint distribution $p(x_a, x_b)$.

8.27 An example is given by

	$x = 0$	$x = 1$	$x = 2$
$y = 0$	0.0	0.1	0.2
$y = 1$	0.0	0.1	0.2
$y = 2$	0.3	0.1	0.0

for which $\widehat{x} = 2$ and $\widehat{y} = 2$.

8.28 If a graph has one or more cycles, there exists at least one set of nodes and edges such that, starting from an arbitrary node in the set, we can visit all the nodes in the set and return to the starting node, without traversing any edge more than once.

Consider one particular such cycle. When one of the nodes n_1 in the cycle sends a message to one of its neighbours n_2 in the cycle, this causes a pending messages on the edge to the next node n_3 in that cycle. Thus sending a pending message along an edge in the cycle always generates a pending message on the next edge in that cycle. Since this is true for every node in the cycle it follows that there will always exist at least one pending message in the graph.

8.29 We show this by induction over the number of nodes in the tree-structured factor graph.

First consider a graph with two nodes, in which case only two messages will be sent across the single edge, one in each direction. None of these messages will induce any pending messages and so the algorithm terminates.

We then assume that for a factor graph with N nodes, there will be no pending messages after a finite number of messages have been sent. Given such a graph, we can construct a new graph with $N + 1$ nodes by adding a new node. This new node will have a single edge to the original graph (since the graph must remain a tree) and so if this new node receives a message on this edge, it will induce no pending messages. A message sent from the new node will trigger propagation of messages in the original graph with N nodes, but by assumption, after a finite number of messages have been sent, there will be no pending messages and the algorithm will terminate.

Chapter 9 Mixture Models and EM

9.1 Since both the E- and the M-step minimise the distortion measure (9.1), the algorithm will never change from a particular assignment of data points to prototypes, unless the new assignment has a lower value for (9.1).

Since there is a finite number of possible assignments, each with a corresponding unique minimum of (9.1) w.r.t. the prototypes, $\{\boldsymbol{\mu}_k\}$, the K-means algorithm will converge after a finite number of steps, when no re-assignment of data points to prototypes will result in a decrease of (9.1). When no-reassignment takes place, there also will not be any change in $\{\boldsymbol{\mu}_k\}$.

9.2 Taking the derivative of (9.1), which in this case only involves $\mathbf{x}_n$, w.r.t. $\boldsymbol{\mu}_k$, we get

$$\frac{\partial J}{\partial \mu_k} = -2r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = z(\boldsymbol{\mu}_k).$$

Substituting this into (2.129), with $\boldsymbol{\mu}_k$ replacing θ, we get

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \eta_n(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})$$

where by (9.2), $\boldsymbol{\mu}_k^{\text{old}}$ will be the prototype nearest to $\mathbf{x}_n$ and the factor of 2 has been absorbed into η_n.

9.3 From (9.10) and (9.11), we have

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

Exploiting the 1-of-K representation for $\mathbf{z}$, we can re-write the r.h.s. as

$$\sum_{j=1}^{K} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{I_{kj}} = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

where $I_{kj} = 1$ if $k = j$ and 0 otherwise.

9.4 From Bayes' theorem we have

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X}}.$$

To maximize this w.r.t. $\boldsymbol{\theta}$, we only need to consider the numerator on the r.h.s. and we shall find it more convenient to operate with the logarithm of this expression,

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \tag{243}$$

where we recognize the first term as the l.h.s. of (9.29). Thus we follow the steps in Section 9.3 in dealing with the latent variables, $\mathbf{Z}$. Note that the second term in

(243) does not involve $\mathbf{Z}$ and will not affect the corresponding E-step, which hence gives (9.30). In the M-step, however, we are maximizing over $\boldsymbol{\theta}$ and so we need to include the second term of (243), yielding

$$\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta}).$$

9.5 Consider any two of the latent variable nodes, which we denote $\mathbf{z}_l$ and $\mathbf{z}_m$. We wish to determine whether these variables are independent, conditioned on the observed data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and on the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$. To do this we consider every possible path from $\mathbf{z}_l$ to $\mathbf{z}_m$. The plate denotes that there are N separate copies of the notes $\mathbf{z}_n$ and $\mathbf{x}_n$. Thus the only paths which connect $\mathbf{z}_l$ and $\mathbf{z}_m$ are those which go via one of the parameter nodes $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ or $\boldsymbol{\pi}$. Since we are conditioning on these parameters they represent observed nodes. Furthermore, any path through one of these parameter nodes must be tail-to-tail at the parameter node, and hence all such paths are blocked. Thus $\mathbf{z}_l$ and $\mathbf{z}_m$ are independent, and since this is true for any pair of such nodes it follows that the posterior distribution factorizes over the data set.

9.6 In this case, the expected complete-data log likelihood function becomes

$$\mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \left\{\ln \pi_k + \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right)\right\}$$

where $\gamma(z_{nk})$ is defined in (9.16). Differentiating this w.r.t. $\boldsymbol{\Sigma}^{-1}$, using (C.24) and (C.28), we get

$$\frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^{\mathrm{T}}$$

where we have also used that $\sum_{k=1}^{K} \gamma(z_{nk}) = 1$ for all n. Setting this equal to zero and rearranging, we obtain

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}.$$

9.7 Consider first the optimization with respect to the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. For this we can ignore the terms in (9.36) which depend on $\ln \pi_k$. We note that, for each data point n, the quantities z_{nk} are all zero except for a particular element which equals one. We can therefore partition the data set into K groups, denoted $\mathbf{X}_k$, such that all the data points $\mathbf{x}_n$ assigned to component k are in group $\mathbf{X}_k$. The complete-data log likelihood function can then be written

$$\ln p\left(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\right) = \sum_{k=1}^{K}\left\{\sum_{n \in \mathbf{X}_k} \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}.$$

This represents the sum of K independent terms, one for each component in the mixture. When we maximize this term with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ we will simply be fitting the k^{th} component to the data set $\mathbf{X}_k$, for which we will obtain the usual maximum likelihood results for a single Gaussian, as discussed in Chapter 2.

For the mixing coefficients we need only consider the terms in $\ln \pi_k$ in (9.36), but we must introduce a Lagrange multiplier to handle the constraint $\sum_k \pi_k = 1$. Thus we maximize

$$\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

which gives

$$0 = \sum_{n=1}^{N} \frac{z_{nk}}{\pi_k} + \lambda.$$

Multiplying through by π_k and summing over k we obtain $\lambda = -N$, from which we have

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk} = \frac{N_k}{N}$$

where N_k is the number of data points in group $\mathbf{X}_k$.

9.8 Using (2.43), we can write the r.h.s. of (9.40) as

$$-\frac{1}{2} \sum_{n=1}^{N}\sum_{j=1}^{K} \gamma(z_{nj})(\mathbf{x}_n - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_j) + \text{const.},$$

where 'const.' summarizes terms independent of $\boldsymbol{\mu}_j$ (for all j). Taking the derivative of this w.r.t. $\boldsymbol{\mu}_k$, we get

$$-\sum_{n=1}^{N} \gamma(z_{nk}) \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \boldsymbol{\Sigma}^{-1} \mathbf{x}_n \right),$$

and setting this to zero and rearranging, we obtain (9.17).

9.9 If we differentiate (9.40) w.r.t. $\boldsymbol{\Sigma}_k^{-1}$, while keeping the $\gamma(z_{nk})$ fixed, we get

$$\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \mathbb{E}_{\mathbf{Z}} \left[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \right] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \frac{1}{2} \left(\boldsymbol{\Sigma}_k - (x_n - \boldsymbol{\mu}_k)(x_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \right)$$

where we have used (C.28). Setting this equal to zero and rearranging, we obtain (9.19).

Appendix E For π_k, we add a Lagrange multiplier term to (9.40) to enforce the constraint

$$\sum_{k=1}^{K} \pi_k = 1$$

yielding

$$\mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})\right] + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).$$

Differentiating this w.r.t. π_k, we get

$$\sum_{n=1}^{N}\gamma(z_{nk})\frac{1}{\pi_k} + \lambda = \frac{N_k}{\pi_k} + \lambda$$

where we have used (9.18). Setting this equal to zero and rearranging, we get

$$N_k = -\pi_k\lambda.$$

Summing both sides over k, making use of (9.9), we see that $-\lambda = N$ and thus

$$\pi_k = \frac{N_k}{N}.$$

9.10 For the mixture model the joint distribution can be written

$$p(\mathbf{x}_a, \mathbf{x}_b) = \sum_{k=1}^{K}\pi_k p(\mathbf{x}_a, \mathbf{x}_b|k).$$

We can find the conditional density $p(\mathbf{x}_b|\mathbf{x}_a)$ by making use of the relation

$$p(\mathbf{x}_b|\mathbf{x}_a) = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{p(\mathbf{x}_a)}.$$

For mixture model the marginal density of $\mathbf{x}_a$ is given by

$$p(\mathbf{x}_a) = \sum_{k=1}^{K}\pi_k p(\mathbf{x}_a|k)$$

where

$$p(\mathbf{x}_a|k) = \int p(\mathbf{x}_a, \mathbf{x}_b|k)\,\mathrm{d}\mathbf{x}_b.$$

Thus we can write the conditional density in the form

$$p(\mathbf{x}_b|\mathbf{x}_a) = \frac{\displaystyle\sum_{k=1}^{K}\pi_k p(\mathbf{x}_a, \mathbf{x}_b|k)}{\displaystyle\sum_{j=1}^{K}\pi_j p(\mathbf{x}_a|j)}.$$

Now we decompose the numerator using

$$p(\mathbf{x}_a, \mathbf{x}_b|k) = p(\mathbf{x}_b|\mathbf{x}_a, k)p(\mathbf{x}_a|k)$$

which allows us finally to write the conditional density as a mixture model of the form

$$p(\mathbf{x}_b|\mathbf{x}_a) = \sum_{k=1}^{K} \lambda_k p(\mathbf{x}_b|\mathbf{x}_a, k) \tag{244}$$

where the mixture coefficients are given by

$$\lambda_k \equiv p(k|\mathbf{x}_a) = \frac{\pi_k p(\mathbf{x}_a|k)}{\sum_j \pi_j p(\mathbf{x}_a|j)} \tag{245}$$

and $p(\mathbf{x}_b|\mathbf{x}_a, k)$ is the conditional for component k.

9.11 As discussed in Section 9.3.2, $\gamma(z_{nk}) \to r_{nk}$ as $\epsilon \to 0$. $\boldsymbol{\Sigma}_k = \epsilon\mathbf{I}$ for all k and are no longer free parameters. π_k will equal the proportion of data points assigned to cluster k and assuming reasonable initialization of $\boldsymbol{\pi}$ and $\{\boldsymbol{\mu}_k\}$, π_k will remain strictly positive. In this situation, we can maximize (9.40) w.r.t. $\{\boldsymbol{\mu}_k\}$ independently of $\boldsymbol{\pi}$, leaving us with

$$\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \epsilon\mathbf{I}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(-\frac{1}{2\epsilon}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2\right) + \text{const.}$$

which equal the negative of (9.1) upto a scaling factor (which is independent of $\{\boldsymbol{\mu}_k\}$).

9.12 Since the expectation of a sum is the sum of the expectations we have

$$\mathbb{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \mathbb{E}_k[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$$

where $\mathbb{E}_k[\mathbf{x}]$ denotes the expectation of $\mathbf{x}$ under the distribution $p(\mathbf{x}|k)$. To find the covariance we use the general relation

$$\text{cov}[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}}$$

to give

$$\begin{aligned}
\text{cov}[\mathbf{x}] &= \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}} \\
&= \sum_{k=1}^{K} \pi_k \mathbb{E}_k[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}} \\
&= \sum_{k=1}^{K} \pi_k \left\{\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}}\right\} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}}.
\end{aligned}$$

9.13 The expectation of $\mathbf{x}$ under the mixture distribution is given by

$$\mathbb{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \mathbb{E}_k[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k.$$

Now we make use of (9.58) and (9.59) to give

$$\begin{aligned}
\mathbb{E}[\mathbf{x}] &= \sum_{k=1}^{K} \pi_k \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\
&= \sum_{n=1}^{N} \mathbf{x}_n \frac{1}{N} \sum_{k=1}^{K} \gamma(z_{nk}) \\
&= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \\
&= \overline{\mathbf{x}}
\end{aligned}$$

where we have used $\pi_k = N_k/N$, and the fact that $\gamma(z_{nk})$ are posterior probabilities and hence $\sum_k \gamma(z_{nk}) = 1$.

Now suppose we initialize a mixture of Bernoulli distributions by setting the means to a common value $\boldsymbol{\mu}_k = \widehat{\boldsymbol{\mu}}$ for $k = 1, \ldots, K$ and then run the EM algorithm. In the E-step we first compute the responsibilities which will be given by

$$\gamma(z_{nk}) = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)} = \frac{\pi_k}{\sum_{j=1}^{K} \pi_j} = \pi_k$$

and are therefore independent of n. In the subsequent M-step the revised means are given by

$$\begin{aligned}
\boldsymbol{\mu}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\
&= \frac{1}{N_k} \pi_k \sum_{n=1}^{N} \mathbf{x}_n \\
&= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \\
&= \overline{\mathbf{x}}
\end{aligned}$$

where again we have made use of $\pi_k = N_k/N$. Note that since these are again the same for all k it follows from the previous discussion that the responsibilities on the

next E-step will again be given by $\gamma(z_{nk}) = \pi_k$ and hence will be unchanged. The revised mixing coefficients are given by

$$\frac{1}{N}\sum_{n=1}^{N}\gamma(z_{nk}) = \pi_k$$

and so are also unchanged. Thus the EM algorithm has converged and no further changes will take place with subsequent E and M steps. Note that this is a degenerate solution in which all of the components of the mixture are identical, and so this distribution is equivalent to a single multivariate Bernoulli distribution.

9.14 Forming the product of (9.52) and (9.53), we get

$$\prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k} \prod_{j=1}^{K} \pi_j^{z_j} = \prod_{k=1}^{K} \left(p(\mathbf{x}|\boldsymbol{\mu}_k)\pi_k\right)^{z_k}.$$

If we marginalize this over $\mathbf{z}$, we get

$$\begin{aligned}\sum_{\mathbf{z}} \prod_{k=1}^{K} \left(p(\mathbf{x}|\boldsymbol{\mu}_k)\pi_k\right)^{z_k} &= \sum_{j=1}^{K}\prod_{k=1}^{K} \left(p(\mathbf{x}|\boldsymbol{\mu}_k)\pi_k\right)^{I_{jk}} \\ &= \sum_{j=1}^{K} \pi_j p(\mathbf{x}|\boldsymbol{\mu}_j)\end{aligned}$$

where we have exploited the 1-of-K coding scheme used for $\mathbf{z}$.

9.15 This is easily shown by calculating the derivatives of (9.55), setting them to zero and solve for μ_{ki}. Using standard derivatives, we get

$$\begin{aligned}\frac{\partial}{\partial \mu_{ki}}\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] &= \sum_{n=1}^{N}\gamma(z_{nk})\left(\frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}\right) \\ &= \frac{\sum_n \gamma(z_{nk})x_{ni} - \sum_n \gamma(z_{nk})\mu_{ki}}{\mu_{ki}(1-\mu_{ki})}.\end{aligned}$$

Setting this to zero and solving for μ_{ki}, we get

$$\mu_{ki} = \frac{\sum_n \gamma(z_{nk})x_{ni}}{\sum_n \gamma(z_{nk})},$$

which equals (9.59) when written in vector form.

9.16 This is identical with the maximization w.r.t. π_k in the Gaussian mixture model, detailed in the second half of Solution 9.9.

9.17 This follows directly from the equation for the incomplete log-likelihood, (9.51). The largest value that the argument to the logarithm on the r.h.s. of (9.51) can have is 1, since $\forall n, k : 0 \leqslant p(\mathbf{x}_n|\boldsymbol{\mu}_k) \leqslant 1$, $0 \leqslant \pi_k \leqslant 1$ and $\sum_k^K \pi_k = 1$. Therefore, the maximum value for $\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\pi})$ equals 0.

9.18 From Solution 9.4, which dealt with MAP estimation for a general mixture model, we know that the E-step will remain unchanged. In the M-step we maximize

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) + \ln p(\boldsymbol{\theta})$$

which in the case of the given model becomes,

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln \pi_k + \sum_{i=1}^{D}\left[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\right]\right\}$$
$$+\sum_{j=1}^{K}\sum_{i'=1}^{D}\left\{(a_j-1)\ln \mu_{ji'} + (b_j-1)\ln(1-\mu_{ji'})\right\} + \sum_{l=1}^{K}(\alpha_l - 1)\ln \pi_l \quad (246)$$

where we have used (9.55), (2.13) and (2.38), and we have dropped terms independent of $\{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\pi}$. Note that we have assumed that each parameter μ_{ki} has the same prior for each i, but this can differ for different components k.

Differentiating (246) w.r.t. μ_{ki} yields

$$\sum_{n=1}^{N}\gamma(z_{nk})\left\{\frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}\right\} + \frac{a_k}{\mu_{ki}} - \frac{1-b_k}{1-\mu_{ki}}$$
$$= \frac{N_k\overline{x}_{ki} + a - 1}{\mu_{ki}} - \frac{N_k - N_k\overline{x}_{ki} + b - 1}{1-\mu_{ki}}$$

where N_k is given by (9.57) and $\overline{x}_{ki}$ is the i^{th} element of $\overline{\mathbf{x}}$ defined in (9.58). Setting this equal to zero and rearranging, we get

$$\mu_{ki} = \frac{N_k\overline{x}_{ki} + a - 1}{N_k + a - 1 + b - 1}. \quad (247)$$

Note that if $a_k = b_k = 1$ for all k, this reduces to the standard maximum likelihood result. Also, as N becomes large, (247) will approach the maximum likelihood result.

When maximizing w.r.t. π_k, we need to enforce the constraint $\sum_k \pi_k = 1$, which we do by adding a Lagrange multiplier term to (246). Dropping terms independent of $\boldsymbol{\pi}$ we are left with

Appendix E

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\ln \pi_k + \sum_{l=1}^{K}(\alpha_l - 1)\ln \pi_l + \lambda\left(\sum_{j=1}^{K}\pi_j - 1\right).$$

Differentiating this w.r.t. π_k, we get

$$\frac{N_k + \alpha_k - 1}{\pi_k} + \lambda$$

and setting this equal to zero and rearranging, we have

$$N_k + \alpha_k - 1 = -\lambda \pi_k.$$

Summing both sides over k, using $\sum_k \pi_k = 1$, we see that $-\lambda = N + \alpha_0 - K$, where α_0 is given by (2.39), and thus

$$\pi_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}. \tag{248}$$

Also in this case, if $\alpha_k = 1$ for all k, we recover the maximum likelihood result exactly. Similarly, as N gets large, (248) will approach the maximum likelihood result.

9.19 As usual we introduce a latent variable $\mathbf{z}_n$ corresponding to each observation. The conditional distribution of the observed data set, given the latent variables, is then

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}_k)^{z_{nk}}.$$

Similarly, the distribution of the latent variables is given by

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \pi_k^{z_{nk}}.$$

The expected value of the complete-data log likelihood function is given by

$$\sum_{n=1}^{N} \sum_{k=1}^{k} \gamma(z_{nk}) \left\{ \ln \pi_k + \sum_{i=1}^{D} \sum_{j=1}^{M} x_{nij} \ln \mu_{kij} \right\}$$

where as usual we have defined responsibilities given by

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{M} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)}.$$

These represent the E-step equations.

To derive the M-step equations we add to the expected complete-data log likelihood function a set of Lagrange multiplier terms given by

$$\lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \sum_{k=1}^{K} \sum_{i=1}^{D} \eta_{ki} \left(\sum_{j=1}^{M} \mu_{kij} - 1 \right)$$

to enforce the constraint $\sum_k \pi_k = 1$ as well as the set of constraints

$$\sum_{j=1}^{M} \mu_{kij} = 1$$

for all values of i and k. Maximizing with respect to the mixing coefficients π_k, and eliminating the Lagrange multiplier λ in the usual way, we obtain

$$\pi_k = \frac{N_k}{N}$$

where we have defined

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

Similarly maximizing with respect to the parameters μ_{kij}, and again eliminating the Lagrange multipliers, we obtain

$$\mu_{kij} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_{nij}.$$

This is an intuitively reasonable result which says that the value of μ_{kij} for component k is given by the fraction of those counts assigned to component k which have non-zero values of the corresponding elements i and j.

9.20 If we take the derivatives of (9.62) w.r.t. α, we get

$$\frac{\partial}{\partial \alpha} \mathbb{E}\left[\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta)\right] = \frac{M}{2}\frac{1}{\alpha} - \frac{1}{2}\mathbb{E}\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right].$$

Setting this equal to zero and re-arranging, we obtain (9.63).

9.21 Taking the derivative of (9.62) w.r.t. β, we obtain

$$\frac{\partial}{\partial \beta} \mathbb{E}\left[\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta)\right] = \frac{N}{2}\frac{1}{\beta} - \frac{1}{2}\sum_{n=1}^{N} \mathbb{E}\left[(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)^2\right]. \tag{249}$$

From (3.49)–(3.51), we see that

$$\begin{aligned}
\mathbb{E}\left[(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)^2\right] &= \mathbb{E}\left[t_n^2 - 2t_n\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n + \mathrm{Tr}[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{w}\mathbf{w}^{\mathrm{T}}]\right] \\
&= t_n^2 - 2t_n\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}_n + \mathrm{Tr}\left[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}(\mathbf{m}_N\mathbf{m}_N^{\mathrm{T}} + \mathbf{S}_N)\right] \\
&= (t_n - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}_n)^2 + \mathrm{Tr}\left[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{S}_N\right].
\end{aligned}$$

Substituting this into (249) and rearranging, we obtain

$$\frac{1}{\beta} = \frac{1}{N}\left(\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \mathrm{Tr}\left[\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{S}_N\right]\right).$$

9.22 **NOTE**: In PRML, a pair of braces is missing from (9.66), which should read

$$\mathbb{E}_{\mathbf{w}}\left[\ln\left\{p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\alpha})\right\}\right].$$

Moreover $\mathbf{m}_N$ should be $\mathbf{m}$ in the numerator on the r.h.s. of (9.68).

Using (7.76)–(7.83) and associated definitions, we can rewrite (9.66) as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}}\left[\ln\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}\right)+\ln\mathcal{N}\left(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1}\right)\right]\\
&\quad=\frac{1}{2}\,\mathbb{E}_{\mathbf{w}}\left[N\ln\beta-\beta\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\|^2+\sum_{i=1}^{M}\ln\alpha_i-\mathrm{Tr}\left[\mathbf{A}\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\right]+\mathrm{const}\\
&\quad=\frac{1}{2}\Bigg(N\ln\beta-\beta\left(\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}\|^2+\mathrm{Tr}[\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Sigma}]\right)\\
&\qquad\qquad+\sum_{i=1}^{M}\ln\alpha_i-\mathrm{Tr}[\mathbf{A}(\mathbf{m}\mathbf{m}^{\mathrm{T}}+\boldsymbol{\Sigma})]\Bigg)+\mathrm{const}.
\end{aligned}\tag{250}$$

Differentiating this w.r.t. α_i, using (C.23), and setting the result equal to zero, we get

$$\frac{1}{2}\frac{1}{\alpha_i}-\frac{1}{2}\left(m_i^2+\Sigma_{ii}\right)=0$$

which we can rearrange to obtain (9.67).

Differentiating (250) w.r.t. β and setting the result equal to zero we get

$$\frac{N}{2}\frac{1}{\beta}-\frac{1}{2}\left(\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}\|^2+\mathrm{Tr}\left[\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Sigma}\right]\right)=0.\tag{251}$$

Using (7.83), (C.6) and (C.7) together with the fact that $\mathbf{A}$ is diagonal, we can rewrite $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Sigma}$ as follows:

$$\begin{aligned}
\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Sigma}&=\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{A}^{-1}\left(\mathbf{I}+\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{A}^{-1}\right)^{-1}\\
&=\boldsymbol{\Phi}^{\mathrm{T}}\left(\mathbf{I}+\beta\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\right)^{-1}\boldsymbol{\Phi}\mathbf{A}^{-1}\\
&=\beta\left(\mathbf{I}-\mathbf{I}+\boldsymbol{\Phi}^{\mathrm{T}}\left(\beta^{-1}\mathbf{I}+\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\right)^{-1}\boldsymbol{\Phi}\mathbf{A}^{-1}\right)\\
&=\beta\left(\mathbf{I}-\mathbf{A}\left(\mathbf{A}^{-1}+\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\left(\beta^{-1}\mathbf{I}+\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\right)^{-1}\boldsymbol{\Phi}\mathbf{A}^{-1}\right)\right)\\
&=\beta\left(\mathbf{I}-\mathbf{A}\left(\mathbf{A}+\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\right)=\beta\left(\mathbf{I}-\mathbf{A}\boldsymbol{\Sigma}\right).
\end{aligned}$$

Using this together with (7.89), we obtain (9.68) from (251).

9.23 **NOTE**: In the 1st printing of PRML, the task set in this exercise is to show that the two sets of re-estimation equations are formally equivalent, without any restriction. However, it really should be restricted to stationary points of the objective function.

Considering the case when the optimization has converged, we can start with α_i, as defined by (7.87), and use (7.89) to re-write this as

$$\alpha_i^\star = \frac{1 - \alpha_i^\star \Sigma_{ii}}{m_N^2},$$

where $\alpha_i^\star = \alpha_i^{\text{new}} = \alpha_i$ is the value reached at convergence. We can re-write this as

$$\alpha_i^\star (m_i^2 + \Sigma_{ii}) = 1$$

which is easily re-written as (9.67).

For β, we start from (9.68), which we re-write as

$$\frac{1}{\beta^\star} = \frac{\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2}{N} + \frac{\sum_i \gamma_i}{\beta^\star N}.$$

As in the α-case, $\beta^\star = \beta^{\text{new}} = \beta$ is the value reached at convergence. We can re-write this as

$$\frac{1}{\beta^\star}\left(N - \sum_i \gamma_i\right) = \|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2,$$

which can easily be re-written as (7.88).

9.24 This is analogous to Solution 10.1, with the integrals replaced by sums.

9.25 This follows from the fact that the Kullback-Leibler divergence, $\mathrm{KL}(q\|p)$, is at its minimum, 0, when q and p are identical. This means that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{KL}(q\|p) = \mathbf{0},$$

since $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$. Therefore, if we compute the gradient of both sides of (9.70) w.r.t. $\boldsymbol{\theta}$, the contribution from the second term on the r.h.s. will be $\mathbf{0}$, and so the gradient of the first term must equal that of the l.h.s.

9.26 From (9.18) we get

$$N_k^{\text{old}} = \sum_n \gamma^{\text{old}}(z_{nk}). \tag{252}$$

We get N_k^{new} by recomputing the responsibilities, $\gamma(z_{mk})$, for a specific data point, $\mathbf{x}_m$, yielding

$$N_k^{\text{new}} = \sum_{n \neq m} \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}). \tag{253}$$

Combining this with (252), we get (9.79).

Similarly, from (9.17) we have

$$\boldsymbol{\mu}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}} \sum_n \gamma^{\text{old}}(z_{nk}) \mathbf{x}_n$$

and recomputing the responsibilities, $\gamma(z_{mk})$, we get

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}}\left(\sum_{n\neq m}\gamma^{\text{old}}(z_{nk})\mathbf{x}_n + \gamma^{\text{new}}(z_{mk})\mathbf{x}_m\right)\\
&= \frac{1}{N_k^{\text{new}}}\left(N_k^{\text{old}}\boldsymbol{\mu}_k^{\text{old}} - \gamma^{\text{old}}(z_{mk})\mathbf{x}_m + \gamma^{\text{new}}(z_{mk})\mathbf{x}_m\right)\\
&= \frac{1}{N_k^{\text{new}}}\Bigg(\left(N_k^{\text{new}} - \gamma^{\text{new}}(z_{mk}) + \gamma^{\text{old}}(z_{mk})\right)\boldsymbol{\mu}_k^{\text{old}}\\
&\qquad -\gamma^{\text{old}}(z_{mk})\mathbf{x}_m + \gamma^{\text{new}}(z_{mk})\mathbf{x}_m\Bigg)\\
&= \boldsymbol{\mu}_k^{\text{old}} + \left(\frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}}\right)(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}),
\end{aligned}$$

where we have used (9.79).

9.27 Following the treatment of $\boldsymbol{\mu}_k$ in Solution 9.26, (9.19) gives

$$\boldsymbol{\Sigma}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})^{\mathrm{T}}$$

where N_k^{old} is given by (252). Recomputing the responsibilities $\gamma(z_{mk})$, and using (253), we get

$$\begin{aligned}
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}}\Bigg(\sum_{n\neq m}\gamma^{\text{old}}(z_{nk})\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}}\right)^{\mathrm{T}}\\
&\qquad +\gamma^{\text{new}}(z_{mk})\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}\Bigg)\\
&= \frac{1}{N_k^{\text{new}}}\Big(N_k^{\text{old}}\boldsymbol{\Sigma}_k^{\text{old}} - \gamma^{\text{old}}(z_{mk})\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^{\mathrm{T}}\\
&\qquad +\gamma^{\text{new}}(z_{mk})\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}\Big)\\
&= \boldsymbol{\Sigma}_k^{\text{old}} - \frac{\gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}}\left(\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^{\mathrm{T}} - \boldsymbol{\Sigma}_k^{\text{old}}\right)\\
&\qquad +\frac{\gamma^{\text{new}}(z_{mk})}{N_k^{\text{new}}}\left(\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} - \boldsymbol{\Sigma}_k^{\text{old}}\right)
\end{aligned}$$

where we have also used (9.79).

For π_k, (9.22) gives

$$\pi_k^{\text{old}} = \frac{N_k^{\text{old}}}{N} = \frac{1}{N}\sum_{n=1}^{N}\gamma^{\text{old}}(z_{nk})$$

and thus recomputing $\gamma(z_{nk})$ we get

$$\begin{aligned}\pi_k^{\text{new}} &= \frac{1}{N}\left(\sum_{n\neq m}^{N}\gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk})\right)\\ &= \frac{1}{N}\left(N\pi_k^{\text{old}} - \gamma^{\text{old}}(z_{mk}) + \gamma^{\text{new}}(z_{mk})\right)\\ &= \pi_k^{\text{old}} - \frac{\gamma^{\text{old}}(z_{mk})}{N} + \frac{\gamma^{\text{new}}(z_{mk})}{N}.\end{aligned}$$

Chapter 10 Approximate Inference

10.1 Starting from (10.3), we use the product rule together with (10.4) to get

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z})\ln\left\{\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right\}\mathrm{d}\mathbf{Z}\\ &= \int q(\mathbf{Z})\ln\left\{\frac{p(\mathbf{X}\mid\mathbf{Z})\,p(\mathbf{X})}{q(\mathbf{Z})}\right\}\mathrm{d}\mathbf{Z}\\ &= \int q(\mathbf{Z})\left(\ln\left\{\frac{p(\mathbf{X}\mid\mathbf{Z})}{q(\mathbf{Z})}\right\} + \ln p(\mathbf{X})\right)\mathrm{d}\mathbf{Z}\\ &= -\mathrm{KL}(\,q\parallel p\,) + \ln p(\mathbf{X})\,.\end{aligned}$$

Rearranging this, we immediately get (10.2).

10.2 By substituting $\mathbb{E}[z_1] = m_1 = \mu_1$ and $\mathbb{E}[z_2] = m_2 = \mu_2$ in (10.13) and (10.15), respectively, we see that both equations are satisfied and so this is a solution.

To show that it is indeed the only solution when $p(\mathbf{z})$ is non-singular, we first substitute $\mathbb{E}[z_1] = m_1$ and $\mathbb{E}[z_2] = m_2$ in (10.13) and (10.15), respectively. Next, we substitute the r.h.s. of (10.13) for m_1 in (10.15), yielding

$$\begin{aligned}m_2 &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(m_2 - \mu_2\right) - \mu_1\right)\\ &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2)\end{aligned}$$

which we can rewrite as

$$m_2\left(1 - \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}\right) = \mu_2\left(1 - \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}\right).$$

Thus, unless $\Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12} = 1$, the solution $\mu_2 = m_2$ is unique. If $p(\mathbf{z})$ is non-singular,

$$|\mathbf{\Lambda}| = \Lambda_{11}\Lambda_{22} - \Lambda_{21}\Lambda_{12} \neq 0$$

which we can rewrite as

$$\Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{21}\Lambda_{12} \neq 1$$

as desired. Since $\mu_2 = m_2$ is the unique solution to (10.15), $\mu_1 = m_1$ is the unique solution to (10.13).

10.3 Starting from (10.16) and optimizing w.r.t. $q_j(\mathbf{Z}_j)$, we get

$$\begin{aligned}
\mathrm{KL}(p \,\|\, q) &= -\int p(\mathbf{Z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \right] \mathrm{d}\mathbf{Z} + \text{const.} \\
&= -\int \left(p(\mathbf{Z}) \ln q_j(\mathbf{Z}_j) + p(\mathbf{Z}) \sum_{i \neq j} \ln q_i(\mathbf{Z}_i) \right) \mathrm{d}\mathbf{Z} + \text{const.} \\
&= -\int p(\mathbf{Z}) \ln q_j(\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z} + \text{const.} \\
&= -\int \ln q_j(\mathbf{Z}_j) \left[\int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i \right] \mathrm{d}\mathbf{Z}_j + \text{const.} \\
&= -\int F_j(\mathbf{Z}_j) \ln q_j(\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j + \text{const.},
\end{aligned}$$

where terms independent of $q_j(\mathbf{Z}_j)$ have been absorbed into the constant term and we have defined

$$F_j(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i.$$

We use a Lagrange multiplier to ensure that $q_j(\mathbf{Z}_j)$ integrates to one, yielding

$$-\int F_j(\mathbf{Z}_j) \ln q_j(\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j + \lambda \left(\int q_j(\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - 1 \right).$$

Using the results from Appendix D, we then take the functional derivative of this w.r.t. q_j and set this to zero, to obtain

$$-\frac{F_j(\mathbf{Z}_j)}{q_j(\mathbf{Z}_j)} + \lambda = 0.$$

From this, we see that

$$\lambda q_j(\mathbf{Z}_j) = F_j(\mathbf{Z}_j).$$

Integrating both sides over $\mathbf{Z}_j$, we see that, since $q_j(\mathbf{Z}_j)$ must intgrate to one,

$$\lambda = \int F_j(\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j = \int \left[\int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i \right] \mathrm{d}\mathbf{Z}_j = 1,$$

and thus

$$q_j(\mathbf{Z}_j) = F_j(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i.$$

10.4 The Kullback-Leibler divergence takes the form

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln q(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Substituting the Gaussian for $q(\mathbf{x})$we obtain

$$\begin{aligned}
\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \left\{ -\frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \mathrm{d}\mathbf{x} + \text{const.} \\
&= \frac{1}{2} \left\{ \ln |\boldsymbol{\Sigma}| + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1} \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \right] \right) \right\} + \text{const.} \\
&= \frac{1}{2} \left\{ \ln |\boldsymbol{\Sigma}| + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - 2 \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbb{E}[\mathbf{x}] + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1} \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}} \right] \right) \right\} \\
&\quad + \text{const.} \qquad (254)
\end{aligned}$$

Differentiating this w.r.t. $\boldsymbol{\mu}$, using results from Appendix C, and setting the result to zero, we see that

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]. \qquad (255)$$

Similarly, differentiating (254) w.r.t. $\boldsymbol{\Sigma}^{-1}$, again using results from Appendix C and also making use of (255) and (1.42), we see that

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}} \right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} = \mathrm{cov}[\mathbf{x}].$$

10.5 We assume that $q(\mathbf{Z}) = q(\mathbf{z})q(\boldsymbol{\theta})$ and so we can optimize w.r.t. $q(\mathbf{z})$ and $q(\boldsymbol{\theta})$ independently.

For $q(\mathbf{z})$, this is equivalent to minimizing the Kullback-Leibler divergence, (10.4), which here becomes

$$\mathrm{KL}(\, q \parallel p\,) = -\iint q\,(\boldsymbol{\theta})\, q\,(\mathbf{z}) \ln \frac{p\,(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{X})}{q\,(\mathbf{z})\, q\,(\boldsymbol{\theta})}\, \mathrm{d}\mathbf{z}\, \mathrm{d}\boldsymbol{\theta}.$$

For the particular chosen form of $q(\boldsymbol{\theta})$, this is equivalent to

$$\begin{aligned}
\mathrm{KL}(\, q \parallel p\,) &= -\int q\,(\mathbf{z}) \ln \frac{p\,(\mathbf{z}, \boldsymbol{\theta}_0 \mid \mathbf{X})}{q\,(\mathbf{z})}\, \mathrm{d}\mathbf{z} + \text{const.} \\
&= -\int q\,(\mathbf{z}) \ln \frac{p\,(\mathbf{z} \mid \boldsymbol{\theta}_0, \mathbf{X})\, p\,(\boldsymbol{\theta}_0 \mid \mathbf{X})}{q\,(\mathbf{z})}\, \mathrm{d}\mathbf{z} + \text{const.} \\
&= -\int q\,(\mathbf{z}) \ln \frac{p\,(\mathbf{z} \mid \boldsymbol{\theta}_0, \mathbf{X})}{q\,(\mathbf{z})}\, \mathrm{d}\mathbf{z} + \text{const.},
\end{aligned}$$

where const accumulates all terms independent of $q(\mathbf{z})$. This KL divergence is minimized when $q(\mathbf{z}) = p(\mathbf{z}|\boldsymbol{\theta}_0, \mathbf{X})$, which corresponds exactly to the E-step of the EM algorithm.

To determine $q(\boldsymbol{\theta})$, we consider

$$\begin{aligned}&\int q(\boldsymbol{\theta}) \int q(\mathbf{z}) \ln \frac{p(\mathbf{X}, \boldsymbol{\theta}, \mathbf{z})}{q(\boldsymbol{\theta}) q(\mathbf{z})} \, \mathrm{d}\mathbf{z} \, \mathrm{d}\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \, \mathbb{E}_{q(\mathbf{z})} [\ln p(\mathbf{X}, \boldsymbol{\theta}, \mathbf{z})] \, \mathrm{d}\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} + \text{const.}\end{aligned}$$

where the last term summarizes terms independent of $q(\boldsymbol{\theta})$. Since $q(\boldsymbol{\theta})$ is constrained to be a point density, the contribution from the entropy term (which formally diverges) will be constant and independent of $\boldsymbol{\theta}_0$. Thus, the optimization problem is reduced to maximizing expected complete log posterior distribution

$$\mathbb{E}_{q(\mathbf{z})} [\ln p(\mathbf{X}, \boldsymbol{\theta}_0, \mathbf{z})],$$

w.r.t. $\boldsymbol{\theta}_0$, which is equivalent to the M-step of the EM algorithm.

10.6 We start by rewriting (10.19) as

$$\mathrm{D}(p\|q) = \frac{4}{1-\alpha^2} \left(1 - \int p(x)^{\gamma_p} q(x)^{\gamma_q} \, \mathrm{d}x\right) \tag{256}$$

so that

$$\gamma_p = \frac{1+\alpha}{2} \quad \text{and} \quad \gamma_q = \frac{1-\alpha}{2}. \tag{257}$$

We note that

$$\lim_{\alpha \to 1} \gamma_q = 0 \tag{258}$$

$$\lim_{\alpha \to 1} \gamma_p = 1 \tag{259}$$

$$1 - \gamma_p = \gamma_q. \tag{260}$$

Based on observation (258), we make a Maclaurin expansion of $q(x)^{\gamma_q}$ in γ_q as follows

$$q^{\gamma_q} = \exp(\gamma_q \ln q) = 1 + \gamma_q \ln q + O\left(\gamma_q^2\right) \tag{261}$$

where q is a shorthand notation for $q(x)$. Similarly, based on (259), we make a Taylor expansion of $p(x)^{\gamma_p}$ in γ_p around 1,

$$\begin{aligned}p^{\gamma_p} &= \exp(\gamma_p \ln p) \\ &= p - (1-\gamma_p) p \ln p + O\left((\gamma_p - 1)^2\right) \\ &= p - \gamma_q p \ln p + O\left(\gamma_q^2\right)\end{aligned} \tag{262}$$

where we have used (260) and we have adopted corresponding shorthand notation for $p(x)$.

Using (261) and (262), we can rewrite (256) as

$$\begin{aligned}
&\mathrm{D}(p\|q) \\
&= \frac{4}{1-\alpha^2}\left(1-\int \left[p-\gamma_q p\ln p+O\left(\gamma_q^2\right)\right]\left[1+\gamma_q \ln q+O\left(\gamma_q^2\right)\right]\,\mathrm{d}x\right) \\
&= \frac{4}{1-\alpha^2}\left(1-\int p+\gamma_q\left(p\ln q-p\ln p\right)\,\mathrm{d}x+O\left(\gamma_q^2\right)\right) \qquad (263)
\end{aligned}$$

where $O\left(\gamma_q^2\right)$ account for all higher order terms. From (257) we have

$$\begin{aligned}
\frac{4}{1-\alpha^2}\gamma_q &= \frac{2(1-\alpha)}{1-\alpha^2}=\frac{2}{(1+\alpha)} \\
\frac{4}{1-\alpha^2}\gamma_q^2 &= \frac{(1-\alpha)^2}{1-\alpha^2}=\frac{1-\alpha}{(1+\alpha)}
\end{aligned}$$

and thus

$$\begin{aligned}
\lim_{\alpha\to 1}\frac{4}{1-\alpha^2}\gamma_q &= 1 \\
\lim_{\alpha\to 1}\frac{4}{1-\alpha^2}\gamma_q^2 &= 0.
\end{aligned}$$

Using these results together with (259), and (263), we see that

$$\lim_{\alpha\to 1}\mathrm{D}(p\|q)=-\int p(\ln q-\ln p)\,\mathrm{d}x=\mathrm{KL}(p\|q).$$

The proof that $\alpha\to -1$ yields $\mathrm{KL}(q\|p)$ is analogous.

10.7 **NOTE**: See note in Solution 10.9.

We take the μ-dependent term from the last line of (10.25) as our starting point. We can rewrite this as follows

$$\begin{aligned}
&-\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0\left(\mu-\mu_0\right)^2+\sum_{n=1}^{N}\left(x_n-\mu\right)^2\right\} \\
&= -\frac{\mathbb{E}[\tau]}{2}\left\{\left(\lambda_0+N\right)\mu^2+\sum_{n=1}^{N}x_n^2-2\mu\left(\lambda_0\mu_0+N\overline{x}\right)\right\} \\
&= -\frac{\mathbb{E}[\tau]}{2}\left\{\left(\lambda_0+N\right)\left(\mu-\frac{\lambda_0\mu_0+N\overline{x}}{\lambda_0+N}\right)^2+\sum_{n=1}^{N}x_n^2-\frac{\left(\lambda_0 m u_0+N\overline{x}\right)^2}{\lambda_0+N}\right\}
\end{aligned}$$

where in the last step we have completed the square over μ. The last two terms are indepenedent of μ and hence can be dropped. Taking the exponential of the remainder, we obtain an unnormalized Gaussian with mean and precision given by (10.26) and (10.27), respectively.

For the posterior over τ, we take the last two lines of (10.28) as our starting point. Gathering terms that multiply τ and $\ln \tau$ into two groups, we can rewrite this as

$$\left(a_0 + \frac{N+1}{2} - 1\right) \ln \tau - \left(b_0 + \frac{1}{2}\mathbb{E}\left[\sum_{n=1}^{N}(x-\mu)^2 + \lambda_0\left(\mu - \mu_0\right)\right]\right)\tau + \text{const.}$$

Taking the exponential of this we get an unnormalized Gamma distribution with shape and inverse scale parameters given by (10.29) and (10.30), respectively.

10.8 **NOTE**: See note in Solution 10.9.

If we substitute the r.h.s. of (10.29) and (10.30) for a and b, respectively, in (B.27) and (B.28), we get

$$\begin{aligned}
\mathbb{E}[\tau] &= \frac{2a_0 + N + 1}{2b_0 + \mathbb{E}\left[\lambda_0(\mu - \mu_0) + \sum_{n=1}^{N}(x_n - \mu)^2\right]} \\
\text{var}[\tau] &= \frac{2a_0 + N + 1}{2\left(b_0 + \frac{1}{2}\mathbb{E}\left[\lambda_0(\mu - \mu_0) + \sum_{n=1}^{N}(x_n - \mu)^2\right]\right)^2} \\
&= \frac{\mathbb{E}[\tau]}{b_0 + \frac{1}{2}\mathbb{E}\left[\lambda_0(\mu - \mu_0) + \sum_{n=1}^{N}(x_n - \mu)^2\right]}
\end{aligned}$$

From this we see directly that

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}[\tau] &= \frac{N}{\mathbb{E}\left[\sum_{n=1}^{N}(x_n - \mu)^2\right]} \\
\lim_{N\to\infty} \text{var}[\tau] &= 0
\end{aligned}$$

as long as the data set is not singular.

10.9 **NOTE**: In the 1^{st} printing of PRML, an extra term of $1/2$ should be added to the r.h.s. of (10.29), with consequential changes to (10.31) and (10.33), which should read

$$\frac{1}{\mathbb{E}[\tau]} = \mathbb{E}\left[\frac{1}{N+1}\sum_{n=1}^{N}(x_n - \mu)^2\right] = \frac{N}{N+1}\left(\overline{x^2} - 2\overline{x}\mathbb{E}[\mu] + \mathbb{E}[\mu^2]\right)$$

and

$$\frac{1}{\mathbb{E}[\tau]} = (\overline{x^2} - \overline{x}^2) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2$$

respectively.

Assuming $a_0 = b_0 = \lambda_0 = 0$, (10.29), (10.30) and (B.27) give

$$\begin{aligned}\frac{1}{\mathbb{E}[\tau]} &= \frac{1}{N+1}\sum_{n=1}^{N}(x_n - \mu)^2 \\ &= \frac{1}{N+1}\sum_{n=1}^{N}\left(x_n^2 - 2x_n\mu + \mu^2\right)\end{aligned}$$

Taking the expectation of this under $q(\mu)$, making use of (10.32), we get

$$\begin{aligned}\frac{1}{\mathbb{E}[\tau]} &= \frac{1}{N+1}\sum_{n=1}^{N}\left(x_n^2 - 2x_n\overline{x} + \overline{x}^2 + \frac{1}{N\mathbb{E}[\tau]}\right) \\ &= \frac{N}{N+1}\left(\frac{1}{N\mathbb{E}[\tau]} - \overline{x}^2 + \frac{1}{N}\sum_{n=1}^{N}x_n^2\right)\end{aligned}$$

which we can rearrange to obtain (10.33).

10.10 **NOTE**: In the 1st printing of PRML, there are errors that affect this exercise. $\mathcal{L}_m$ used in (10.34) and (10.35) should really be $\mathcal{L}$, whereas $\mathcal{L}_m$ used in (10.36) is given in Solution 10.11 below.

This completely analogous to Solution 10.1. Starting from (10.35), we can use the product rule to get,

$$\begin{aligned}\mathcal{L} &= \sum_m\sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m)\ln\left\{\frac{p(\mathbf{Z},\mathbf{X},m)}{q(\mathbf{Z}|m)\,q(m)}\right\} \\ &= \sum_m\sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m)\ln\left\{\frac{p(\mathbf{Z},m|\mathbf{X})\,p(\mathbf{X})}{q(\mathbf{Z}|m)\,q(m)}\right\} \\ &= \sum_m\sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m)\ln\left\{\frac{p(\mathbf{Z},m|\mathbf{X})}{q(\mathbf{Z}|m)\,q(m)}\right\} + \ln p(\mathbf{X}).\end{aligned}$$

Rearranging this, we obtain (10.34).

10.11 **NOTE**: Consult note preceding Solution 10.10 for some relevant corrections.

We start by rewriting the lower bound as follows

$$\begin{aligned}
\mathcal{L} &= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, \mathbf{X}, m)}{q(\mathbf{Z}|m)q(m)} \right\} \\
&= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \left\{ \ln p(\mathbf{Z}, \mathbf{X}|m) + \ln p(m) - \ln q(\mathbf{Z}|m) - \ln q(m) \right\} \\
&= \sum_m q(m) \Bigg(\ln p(m) - \ln q(m) \\
&\qquad + \sum_{\mathbf{Z}} q(\mathbf{Z}|m) \{ \ln p(\mathbf{Z}, \mathbf{X}|m) - \ln q(\mathbf{Z}|m) \} \Bigg) \\
&= \sum_m q(m) \left\{ \ln \left(p(m) \exp\{\mathcal{L}_m\} \right) - \ln q(m) \right\}, \qquad (264)
\end{aligned}$$

where

$$\mathcal{L}_m = \sum_{\mathbf{Z}} q(\mathbf{Z}|m) \ln \left\{ \frac{p(\mathbf{Z}, \mathbf{X}|m)}{q(\mathbf{Z}|m)} \right\}.$$

We recognize (264) as the negative KL divergence between $q(m)$ and the (not necessarily normalized) distribution $p(m) \exp\{\mathcal{L}_m\}$. This will be maximized when the KL divergence is minimized, which will be the case when

$$q(m) \propto p(m) \exp\{\mathcal{L}_m\}.$$

10.12 This derivation is given in detail in Section 10.2.1, starting with the paragraph containing (10.43) (page 476) and ending with (10.49).

10.13 In order to derive the optimal solution for $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ we start with the result (10.54) and keep only those term which depend on $\boldsymbol{\mu}_k$ or $\boldsymbol{\Lambda}_k$ to give

$$\begin{aligned}
\ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &= \ln \mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}\right) + \ln \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0) \\
&\qquad + \sum_{n=1}^N \mathbb{E}[z_{nk}] \ln \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) + \text{const.} \\
&= -\frac{\beta_0}{2} (\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0) + \frac{1}{2} \ln |\boldsymbol{\Lambda}_k| - \frac{1}{2} \mathrm{Tr}\left(\boldsymbol{\Lambda}_k \mathbf{W}_0^{-1}\right) \\
&\quad + \frac{(\nu_0 - D - 1)}{2} \ln |\boldsymbol{\Lambda}_k| - \frac{1}{2} \sum_{n=1}^N \mathbb{E}[z_{nk}] (\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Lambda}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \\
&\quad + \frac{1}{2} \left(\sum_{n=1}^N \mathbb{E}[z_{nk}] \right) \ln |\boldsymbol{\Lambda}_k| + \text{const.} \qquad (265)
\end{aligned}$$

Using the product rule of probability, we can express $\ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ as $\ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ $+ \ln q^\star(\boldsymbol{\Lambda}_k)$. Let us first of all identify the distribution for $\boldsymbol{\mu}_k$. To do this we need only consider terms on the right hand side of (265) which depend on $\boldsymbol{\mu}_k$, giving

$$\begin{aligned}
&\ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) \\
&= -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\left[\beta_0 + \sum_{n=1}^{N}\mathbb{E}[z_{nk}]\right]\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Lambda}_k\left[\beta_0\mathbf{m}_0 + \sum_{n=1}^{N}\mathbb{E}[z_{nk}]\mathbf{x}_n\right] \\
&\quad +\text{const.} \\
&= -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\left[\beta_0 + N_k\right]\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Lambda}_k\left[\beta_0\mathbf{m}_0 + N_k\overline{\mathbf{x}}_k\right] + \text{const.}
\end{aligned}$$

where we have made use of (10.51) and (10.52). Thus we see that $\ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ depends quadratically on $\boldsymbol{\mu}_k$ and hence $q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ is a Gaussian distribution. Completing the square in the usual way allows us to determine the mean and precision of this Gaussian, giving

$$q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \beta_k\boldsymbol{\Lambda}_k) \tag{266}$$

where

$$\begin{aligned}
\beta_k &= \beta_0 + N_k \\
\mathbf{m}_k &= \frac{1}{\beta_k}\left(\beta_0\mathbf{m}_0 + N_k\overline{\mathbf{x}}_k\right).
\end{aligned}$$

Next we determine the form of $q^\star(\boldsymbol{\Lambda}_k)$ by making use of the relation

$$\ln q^\star(\boldsymbol{\Lambda}_k) = \ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) - \ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k).$$

On the right hand side of this relation we substitute for $\ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ using (265), and we substitute for $\ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ using the result (266). Keeping only those terms which depend on $\boldsymbol{\Lambda}_k$ we obtain

$$\begin{aligned}
\ln q^\star(\boldsymbol{\Lambda}_k) = &-\frac{\beta_0}{2}(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0) + \frac{1}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k\mathbf{W}_0^{-1}\right) \\
&+ \frac{(\nu_0 - D - 1)}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) \\
&+ \frac{1}{2}\left(\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\right)\ln|\boldsymbol{\Lambda}_k| + \frac{\beta_k}{2}(\boldsymbol{\mu}_k - \mathbf{m}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_k) \\
&- \frac{1}{2}\ln|\boldsymbol{\Lambda}_k| + \text{const.} \\
= &\frac{(\nu_k - D - 1)}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k\mathbf{W}_k^{-1}\right) + \text{const.}
\end{aligned}$$

Here we have defined

$$\begin{aligned}\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + \beta_0(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + \sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \\ &\quad -\beta_k(\boldsymbol{\mu}_k - \mathbf{m}_k)(\boldsymbol{\mu}_k - \mathbf{m}_k)^{\mathrm{T}} \\ &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\overline{\mathbf{x}}_k - \mathbf{m}_0)(\overline{\mathbf{x}}_k - \mathbf{m}_0)^{\mathrm{T}} \qquad (267) \\ \nu_k &= \nu_0 + \sum_{n=1}^{N}\mathbb{E}[z_{nk}] \\ &= \nu_0 + N_k,\end{aligned}$$

where we have made use of the result

$$\begin{aligned}\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} &= \sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n - \overline{\mathbf{x}}_k)(\mathbf{x}_n - \overline{\mathbf{x}}_k)^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} \\ &= N_k\mathbf{S}_k + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} \qquad (268)\end{aligned}$$

and we have made use of (10.53). Note that the terms involving $\boldsymbol{\mu}_k$ have cancelled out in (267) as we expect since $q^\star(\boldsymbol{\Lambda}_k)$ is independent of $\boldsymbol{\mu}_k$.

Thus we see that $q^\star(\boldsymbol{\Lambda}_k)$ is a Wishart distribution of the form

$$q^\star(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k).$$

10.14 We can express the required expectation as an integration with respect to the variational posterior distribution $q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)q^\star(\boldsymbol{\Lambda}_k)$. Thus we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \\ &= \iint \mathrm{Tr}\left\{\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\right\} q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)q^\star(\boldsymbol{\Lambda}_k)\,\mathrm{d}\boldsymbol{\mu}_k\,\mathrm{d}\boldsymbol{\Lambda}_k.\end{aligned}$$

Next we use the result $q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \beta_k\boldsymbol{\Lambda}_k)$ to perform the integration over $\boldsymbol{\mu}_k$ using the standard expressions for expectations under a Gaussian distribution, giving

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_k] &= \mathbf{m}_k \\ \mathbb{E}[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}}] &= \mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}\end{aligned}$$

from which we obtain the expectation with respect to $\boldsymbol{\mu}_k$ in the form

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\mu}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\right] \\ &= \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \mathbf{x}_n\mathbf{m}_k^{\mathrm{T}} - \mathbf{m}_k\mathbf{x}_n^{\mathrm{T}} + \mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1} \\ &= (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}} + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}.\end{aligned}$$

Finally, taking the expectation with respect to $\mathbf{\Lambda}_k$ we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\mu}_k,\mathbf{\Lambda}_k}\left[(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)\right]\\
&\quad=\int \mathrm{Tr}\left\{\mathbf{\Lambda}_k\left[(\mathbf{x}_n-\mathbf{m}_k)(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}+\beta_k^{-1}\mathbf{\Lambda}_k^{-1}\right]\right\}q^\star(\mathbf{\Lambda}_k)\,\mathrm{d}\mathbf{\Lambda}_k\\
&\quad=\int\left\{(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}\mathbf{\Lambda}_k(\mathbf{x}_n-\mathbf{m}_k)+D\beta_k^{-1}\right\}q^\star(\mathbf{\Lambda}_k)\,\mathrm{d}\mathbf{\Lambda}_k\\
&\quad=D\beta_k^{-1}+\nu_k(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n-\mathbf{m}_k)
\end{aligned}$$

as required. Here we have used $q^\star(\mathbf{\Lambda}_k)=\mathcal{W}(\mathbf{\Lambda}_k|\mathbf{W}_k,\nu_k)$, together with the standard result for the expectation under a Wishart distribution to give $\mathbb{E}[\mathbf{\Lambda}_k]=\nu_k\mathbf{W}_k$.

10.15 By substituting (10.58) into (B.17) and then using (B.24) together with the fact that $\sum_k N_k=N$, we obtain (10.69).

10.16 To derive (10.71) we make use of (10.38) to give

$$\begin{aligned}
&\mathbb{E}[\ln p(D|\mathbf{z},\boldsymbol{\mu},\mathbf{\Lambda})]\\
&\quad=\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left\{\mathbb{E}[\ln|\mathbf{\Lambda}_k|]-\mathbb{E}[(\mathbf{x}_n-\boldsymbol{\mu}_k)\mathbf{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)]-D\ln(2\pi)\right\}.
\end{aligned}$$

We now use $\mathbb{E}[z_{nk}]=r_{nk}$ together with (10.64) and the definition of $\widetilde{\Lambda}_k$ given by (10.65) to give

$$\begin{aligned}
\mathbb{E}[\ln p(D|\mathbf{z},\boldsymbol{\mu},\mathbf{\Lambda})]=\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\Big\{&\ln\widetilde{\Lambda}_k\\
&-D\beta_k^{-1}-\nu_k(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n-\mathbf{m}_k)-D\ln(2\pi)\Big\}.
\end{aligned}$$

Now we use the definitions (10.51) to (10.53) together with the result (268) to give (10.71).

We can derive (10.72) simply by taking the logarithm of $p(\mathbf{z}|\boldsymbol{\pi})$ given by (10.37)

$$\mathbb{E}[\ln p(\mathbf{z}|\boldsymbol{\pi})]=\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\mathbb{E}[\ln\pi_k]$$

and then making use of $\mathbb{E}[z_{nk}]=r_{nk}$ together with the definition of $\widetilde{\pi}_k$ given by (10.65).

10.17 The result (10.73) is obtained by using the definition of $p(\boldsymbol{\pi})$ given by (10.39) together with the definition of $\widetilde{\pi}_k$ given by (10.66).

For the result (10.74) we start with the definition of the prior $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ given by (10.40) to give

$$\mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] =$$
$$\frac{1}{2}\sum_{k=1}^{K}\left\{D\ln\beta_0 - D\ln(2\pi) + \mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] - \beta_0\mathbb{E}[(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)]\right\}$$
$$+ K\ln B(\mathbf{W}_0, \nu_0) + \sum_{k=1}^{K}\left\{\frac{(\nu_0 - D - 1)}{2}\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] - \frac{1}{2}\mathrm{Tr}(\mathbf{W}_0^{-1}\mathbb{E}[\boldsymbol{\Lambda}_k])\right\}.$$

Now consider the term $\mathbb{E}[(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)]$. To evaluate this expression we first perform the expectation with respect to $q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ then the subsequently perform the expectation with respect to $q^\star(\boldsymbol{\Lambda}_k)$. Using the standard results for moments under a Gaussian we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_k] &= \mathbf{m}_k \\ \mathbb{E}[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}}] &= \mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}\end{aligned}$$

and hence

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}_k\boldsymbol{\Lambda}_k}[(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)] &= \mathrm{Tr}\left(\mathbb{E}_{\boldsymbol{\mu}_k\boldsymbol{\Lambda}_k}\left[\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\right]\right) \\ &= \mathrm{Tr}\left(\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\boldsymbol{\Lambda}_k(\beta_k^{-1}\boldsymbol{\Lambda}_k^{-1} + \mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} - \mathbf{m}_0\mathbf{m}_k^{\mathrm{T}} - \mathbf{m}_k\mathbf{m}_0^{\mathrm{T}} + \mathbf{m}_0\mathbf{m}_0^{\mathrm{T}})\right]\right) \\ &= K\beta_k^{-1} + (\mathbf{m}_k - \mathbf{m}_0)^{\mathrm{T}}\mathbb{E}[\boldsymbol{\Lambda}_k](\mathbf{m}_k - \mathbf{m}_0).\end{aligned}$$

Now we use (B.80) to give $\mathbb{E}[\boldsymbol{\Lambda}_k] = \nu_k\mathbf{W}_k$ and $\mathbb{E}[\ln\boldsymbol{\Lambda}_k] = \ln\widetilde{\Lambda}_k$ from (10.65) to give (10.74).

For (10.75) we take use the result (10.48) for $q^\star(\mathbf{z})$ to give

$$\mathbb{E}[\ln q(\mathbf{z})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\ln r_{nk}$$

and using $\mathbb{E}[z_{nk}] = r_{nk}$ we obtain (10.75).

The solution (10.76) for $\mathbb{E}[\ln q(\boldsymbol{\pi})]$ is simply the negative entropy of the corresponding Dirichlet distribution (10.57) and is obtained from (B.22).

Finally, we need the entropy of the Gaussian-Wishart distribution $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. First of all we note that this distribution factorizes into a product of factors $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ and the entropy of the product is the sum of the entropies of the individual terms, as is easily verified. Next we write

$$\ln q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \ln q(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) + \ln q(\boldsymbol{\Lambda}_k).$$

Consider first the quantity $\mathbb{E}[\ln q(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)]$. Taking the expectation first with respect to $\boldsymbol{\mu}_k$ we can make use of the standard result (B.41) for the entropy of a Gaussian to

give

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}_k\boldsymbol{\Lambda}_k}[\ln q(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)] &= \mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\frac{1}{2}\ln|\boldsymbol{\Lambda}_k| + \frac{D}{2}\left(\ln\beta_k - 1 - \ln(2\pi)\right)\right] \\ &= \frac{1}{2}\ln\widetilde{\Lambda}_k + \frac{D}{2}\left(\ln\beta_k - 1 - \ln(2\pi)\right).\end{aligned}$$

The term $\mathbb{E}[\ln q(\boldsymbol{\Lambda}_k)]$ is simply the negative entropy of a Wishart distribution, which we write as $-\mathrm{H}[q(\boldsymbol{\Lambda}_k)]$.

10.18 We start with β_k, which appears in (10.71), (10.74) and (10.77). Using these, we can differentiate (10.70) w.r.t. β_k^{-1}, to get

$$\frac{\partial\mathcal{L}}{\partial\beta_k^{-1}} = \frac{D}{2}\left(-N_k - \beta_0 + \beta_k\right).$$

Setting this equal to zero and rearranging the terms, we obtain (10.60). We then consider $\mathbf{m}_k$, which appears in the quadratic terms of (10.71) and (10.74). Thus differentiation of (10.70) w.r.t. $\mathbf{m}_k$ gives

$$\frac{\partial\mathcal{L}}{\partial\mathbf{m}_k} = -N_k\nu_k\left(\mathbf{W}_k\mathbf{m}_k - \mathbf{W}_k\overline{\mathbf{x}}_k\right) - \beta_0\nu_k\left(\mathbf{W}_k\mathbf{m}_k - \mathbf{W}_k\mathbf{m}_0\right).$$

Setting this equal to zero, using (10.60) and rearranging the terms, we obtain (10.61).

Next we tackle $\{\mathbf{W}_k, \nu_k\}$. Here we need to perform a joint optimization w.r.t. $\mathbf{W}_k$ and ν_k for each $k = 1, \ldots, K$. Like β_k, $\mathbf{W}_k$ and ν_k appear in (10.71), (10.74) and (10.77). Using these, we can rewrite the r.h.s. of (10.70) as

$$\begin{aligned}\frac{1}{2}\sum_k^K\Bigg(&N_k\ln\widetilde{\Lambda}_k - N_k\nu_k\left\{\mathrm{Tr}\left(\mathbf{S}_k\mathbf{W}_k\right) + \mathrm{Tr}\left(\mathbf{W}_k\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)^{\mathrm{T}}\right)\right\} \\ &+ \ln\widetilde{\Lambda}_k - \beta_0\nu_k\left(\mathbf{m}_k - \mathbf{m}_0\right)^{\mathrm{T}}\mathbf{W}_k\left(\mathbf{m}_k - \mathbf{m}_0\right) + \left(\nu_0 - D - 1\right)\ln\widetilde{\Lambda}_k \\ &- \nu_k\mathrm{Tr}\left(\mathbf{W}_0^{-1}\mathbf{W}_k\right) - \ln\widetilde{\Lambda}_k + 2\mathrm{H}\left[q(\boldsymbol{\Lambda}_k\right]\Bigg) \qquad (269)\end{aligned}$$

where we have dropped terms independent of $\{\mathbf{W}_k, \nu_k\}$, $\ln\widetilde{\Lambda}_k$ is given by (10.65),

$$\mathrm{H}\left[q(\boldsymbol{\Lambda}_k\right] = -\ln B(\mathbf{W}_k, \nu_k) - \frac{\nu_k - D - 1}{2}\ln\widetilde{\Lambda}_k\frac{\nu_k D}{2} \qquad (270)$$

where we have used (10.65) and (B.81), and, from (B.79),

$$\ln B(\mathbf{W}_k, \nu_k) = \frac{\nu_k}{2}\ln|\mathbf{W}_k| - \frac{\nu_k D}{2} - \sum_{i=1}^{D}\ln\Gamma\left(\frac{\nu_k + 1 - i}{2}\right). \qquad (271)$$

Restricting attention to a single component, k, and making use of (270), (269) gives

$$\frac{1}{2}\left(N_k + \nu_0 - \nu_k\right)\ln\widetilde{\Lambda}_k - \frac{\nu_k}{2}\mathrm{Tr}\left(\mathbf{W}_k\mathbf{F}_k\right) - \ln B(\mathbf{W}_k, \nu_k) + \frac{\nu_k D}{2} \qquad (272)$$

where

$$\begin{aligned}\mathbf{F}_k &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + N_k\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)^{\mathrm{T}} \\ &\quad + \beta_0\left(\mathbf{m}_k - \mathbf{m}_0\right)\left(\mathbf{m}_k - \mathbf{m}_0\right)^{\mathrm{T}} \\ &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{N_k\beta_0}{N_k+\beta_0}\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)^{\mathrm{T}} \qquad (273)\end{aligned}$$

as we shall show below. Differentiating (272) w.r.t. ν_k, making use of (271) and (10.65), and setting the result to zero, we get

$$\begin{aligned}0 &= \frac{1}{2}\Bigg(\left(N_k+\nu_0-\nu_k\right)\frac{\mathrm{d}\ln\widetilde{\Lambda}_k}{\mathrm{d}\nu_k} - \ln\widetilde{\Lambda}_k - \mathrm{Tr}\left(\mathbf{W}_k\mathbf{F}_k\right) \\ &\quad + \ln|\mathbf{W}_k| + D\ln 2 + \sum_{i=1}^{D}\ln\Gamma\left(\frac{\nu_k+1-i}{2}\right) + D\Bigg) \\ &= \frac{1}{2}\left(\left(N_k+\nu_0-\nu_k\right)\frac{\mathrm{d}\ln\widetilde{\Lambda}_k}{\mathrm{d}\nu_k} - \mathrm{Tr}\left(\mathbf{W}_k\mathbf{F}_k\right) + D\right). \qquad (274)\end{aligned}$$

Similarly, differentiating (272) w.r.t. $\mathbf{W}_k$, making use of (271), (273) and (10.65), and setting the result to zero, we get

$$\begin{aligned}0 &= \frac{1}{2}\left(\left(N_k+\nu_0-\nu_k\right)\mathbf{W}_k^{-1} - \mathbf{F}_k + \mathbf{W}_k^{-1}\right) \\ &= \frac{1}{2}\Bigg(\left(N_k+\nu_0-\nu_k\right)\mathbf{W}_k^{-1} - \mathbf{W}_0^{-1} - N_k\mathbf{S}_k \\ &\quad - \frac{N_k\beta_0}{N_k+\beta_0}\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)^{\mathrm{T}} + \mathbf{W}_k^{-1}\Bigg) \qquad (275)\end{aligned}$$

We see that the simultaneous equations (274) and (275) are satisfied if and only if

$$\begin{aligned}0 &= N_k + \nu_0 - \nu_k \\ 0 &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{N_k\beta_0}{N_k+\beta_0}\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)^{\mathrm{T}} - \mathbf{W}_k^{-1}\end{aligned}$$

from which (10.63) and (10.62) follow, respectively. However, we still have to derive (273). From (10.60) and (10.61), derived above, we have

$$\mathbf{m}_k = \frac{\beta_0\mathbf{m}_0 + N_k\overline{\mathbf{x}}_k}{\beta_0 + N_k}$$

and using this, we get

$$
\begin{aligned}
&N_k\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)^{\mathrm{T}}+\beta_0\left(\mathbf{m}_k-\mathbf{m}_0\right)\left(\mathbf{m}_k-\mathbf{m}_0\right)^{\mathrm{T}}= \\
&\qquad N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}}-N_k\overline{\mathbf{x}}_k\frac{\left(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k\right)^{\mathrm{T}}}{\beta_0+N_k}-\frac{\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k}{\beta_0+N_k}N_k\overline{\mathbf{x}}_k^{\mathrm{T}} \\
&+\frac{N_k(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k)(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k)^{\mathrm{T}}}{(\beta_0+N_k)^2}+\frac{\beta_0(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k)(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k)^{\mathrm{T}}}{(\beta_0+N_k)^2} \\
&\qquad -\beta_0\mathbf{m}_0\frac{\left(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k\right)^{\mathrm{T}}}{\beta_0+N_k}-\frac{\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k}{\beta_0+N_k}\beta_0\mathbf{m}_0^{\mathrm{T}}+\beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}}.
\end{aligned}
$$

We now gather the coefficients of the terms on the r.h.s. as follows.

Coefficients of $\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}}$:

$$
\begin{aligned}
&N_k-\frac{N_k^2}{\beta_0+N_k}-\frac{N_k^2}{\beta_0+N_k}+\frac{N_k^3}{(\beta_0+N_k)^2}+\frac{\beta_0N_k^2}{(\beta_0+N_k)^2} \\
&=\quad N_k-\frac{N_k^2}{\beta_0+N_k}-\frac{N_k^2}{\beta_0+N_k}+\frac{N_k^3}{(\beta_0+N_k)^2}+\frac{\beta_0N_k^2}{(\beta_0+N_k)^2} \\
&=\quad \frac{N_k\beta_0^2+N_k^2\beta_0+N_k^2\beta_0+N_k^3-2(N_k^2\beta_0+N_k^3)+N_k^3+\beta_0N_k^2}{(\beta_0+N_k)^2} \\
&=\quad \frac{N_k\beta_0^2+\beta_0N_k^2}{(\beta_0+N_k)^2}=\frac{N_k\beta_0(\beta_0+N_k)}{(\beta_0+N_k)^2}=\frac{N_k\beta_0}{\beta_0+N_k}
\end{aligned}
$$

Coefficients of $\overline{\mathbf{x}}_k\mathbf{m}_0^{\mathrm{T}}$ and $\mathbf{m}_0\overline{\mathbf{x}}_k^{\mathrm{T}}$ (these are identical):

$$
\begin{aligned}
&-\frac{N_k\beta_0}{\beta_0+N_k}+\frac{\beta_0N_k^2}{(\beta_0+N_k)^2}+\frac{\beta_0^2N_k}{(\beta_0+N_k)^2}-\frac{N_k\beta_0}{\beta_0+N_k} \\
&\qquad =\frac{N_k\beta_0}{\beta_0+N_k}-\frac{2N_k\beta_0}{\beta_0+N_k}=-\frac{N_k\beta_0}{\beta_0+N_k}
\end{aligned}
$$

Coefficients of $\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}}$:

$$
\begin{aligned}
&\frac{N_k\beta_0^2}{(\beta_0+N_k)^2}+\frac{\beta_0^3}{(\beta_0+N_k)^2}-\frac{2\beta_0^2}{\beta_0+N_k}+\beta_0 \\
&=\quad \frac{\beta_0^2(N_k+\beta_0)}{(\beta_0+N_k)^2}-\frac{2\beta_0^2}{\beta_0+N_k}+\beta_0 \\
&=\quad \frac{\beta_0^2-2\beta_0^2+\beta_0^2+N_k\beta_0}{\beta_0+N_k}=\frac{N_k\beta_0}{\beta_0+N_k}
\end{aligned}
$$

Thus

$$
\begin{aligned}
&N_k\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)^{\mathrm{T}}+\beta_0\left(\mathbf{m}_k-\mathbf{m}_0\right)\left(\mathbf{m}_k-\mathbf{m}_0\right)^{\mathrm{T}} \\
&\qquad =\frac{N_k\beta_0}{N_k+\beta_0}\left(\overline{\mathbf{x}}_k-\mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k-\mathbf{m}_0\right)^{\mathrm{T}} \qquad (276)
\end{aligned}
$$

as desired.

Now we turn our attention to $\boldsymbol{\alpha}$, which appear in (10.72) and (10.73), through (10.66), and (10.76). Using these together with (B.23) and (B.24), we can differentiate (10.70) w.r.t. α_k and set it equal to zero, yielding

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \alpha_k} &= [N_k + (\alpha_0 - 1) - (\alpha_k - 1)]\frac{\partial \ln \widehat{\pi}_k}{\partial \alpha_k} - \ln \widehat{\pi}_k - \frac{\partial \ln C(\boldsymbol{\alpha})}{\partial \alpha_k} \\ &= [N_k + (\alpha_0 - 1) - (\alpha_k - 1)]\left\{\psi_1(\alpha_k) - \psi_1(\widehat{\alpha})\frac{\partial \widehat{\alpha}}{\partial \alpha_k}\right\} \\ &\quad + \psi(\widehat{\alpha}) - \psi(\alpha_k) - \psi(\widehat{\alpha})\frac{\partial \widehat{\alpha}}{\partial \alpha_k} + \psi(\alpha_k) \\ &= [N_k + (\alpha_0 - 1) - (\alpha_k - 1)]\left\{\psi_1(\alpha_k) - \psi_1(\widehat{\alpha})\right\} = 0 \end{aligned} \tag{277}$$

where $\psi(\cdot)$ and $\psi_1(\cdot)$ are di- and trigamma functions, respectively. If we assume that $\alpha_0 > 0$, (10.58) must hold for (277) to hold, since the trigamma function is strictly positive and monotoncally decreasing for arguments greater than zero.

Finally, we maximize (10.70) w.r.t. r_{nk}, subject to the constraints $\sum_k r_{nk} = 1$ for all $n = 1, \ldots, N$. Note that r_{nk} not only appears in (10.72) and (10.75), but also in (10.71) through N_k, $\overline{\mathbf{x}}_k$ and $\mathbf{S}_k$, and so we must substitute using (10.51), (10.52) and (10.53), respectively. To simplify subsequent calculations, we start by considering the last two terms inside the braces of (10.71), which we write together as

$$\frac{1}{2}\sum_{k=1}^{K} \nu_k \mathrm{Tr}\left(\mathbf{W}_k \mathbf{Q}_k\right) \tag{278}$$

where, using (10.51), (10.52) and (10.53),

$$\begin{aligned}\mathbf{Q}_k &= \sum_{n=1}^{N} r_{nk}\left(\mathbf{x}_n - \overline{\mathbf{x}}_k\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}_k\right)^{\mathrm{T}} + N_k\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)^{\mathrm{T}} \\ &= \sum_{n=1}^{N} r_{nk}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - 2N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} \\ &\quad + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - N_k\mathbf{m}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - N_k\overline{\mathbf{x}}_k\mathbf{m}_k^{\mathrm{T}} + N_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\ &= \sum_{n=1}^{N} r_{nk}\left(\mathbf{x}_n - \mathbf{m}_k\right)\left(\mathbf{x}_n - \mathbf{m}_k\right)^{\mathrm{T}}. \end{aligned} \tag{279}$$

Using (10.51), (278) and (279), we can now consider all terms in (10.71) which

depend on r_{kn} and add the appropriate Lagrange multiplier terms, yielding

$$\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\left(\ln\widetilde{\Lambda}_k - D\beta_k^{-1}\right) - \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\nu_k\left(\mathbf{x}_n-\mathbf{m}_k\right)^{\mathrm{T}}\mathbf{W}_k\left(\mathbf{x}_n-\mathbf{m}_k\right) + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln\widetilde{\pi}_k - \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln r_{nk} + \sum_{n=1}^{N}\lambda_n\left(1-\sum_{k=1}^{K} r_{nk}\right).$$

Taking the derivative of this w.r.t. r_{kn} and setting it equal to zero we obtain

$$0 = \frac{1}{2}\ln\widetilde{\Lambda}_k - \frac{D}{2\beta_k} - \frac{1}{2}\nu_k\left(\mathbf{x}_n-\mathbf{m}_k\right)^{\mathrm{T}}\mathbf{W}_k\left(\mathbf{x}_n-\mathbf{m}_k\right) + \ln\widetilde{\pi}_k - \ln r_{nk} - 1 - \lambda_n.$$

Moving $\ln r_{nk}$ to the l.h.s. and exponentiating both sides, we see that for each n,

$$r_{nk} \propto \widetilde{\pi}_k\widetilde{\Lambda}_k^{1/2}\exp\left\{-\frac{D}{2\beta_k} - \frac{1}{2}\nu_k\left(\mathbf{x}_n-\mathbf{m}_k\right)^{\mathrm{T}}\mathbf{W}_k\left(\mathbf{x}_n-\mathbf{m}_k\right)\right\}$$

which is in agreement with (10.67); the normalized form is then given by (10.49).

10.19 We start by performing the integration over $\boldsymbol{\pi}$ in (10.80), making use of the result

$$\mathbb{E}[\pi_k] = \frac{\alpha_k}{\widehat{\alpha}}$$

to give

$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K}\frac{\alpha_k}{\widehat{\alpha}}\iint \mathcal{N}(\widehat{\mathbf{x}}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)\,\mathrm{d}\boldsymbol{\mu}_k\,\mathrm{d}\boldsymbol{\Lambda}_k.$$

The variational posterior distribution over $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ is given from (10.59) by

$$q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k,(\beta_k\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k,\nu_k).$$

Using this result we next perform the integration over $\boldsymbol{\mu}_k$. This can be done explicitly by completing the square in the exponential in the usual way, or we can simply appeal to the general result (2.109) and (2.110) for the linear-Gaussian model from Chapter 2 to give

$$\int \mathcal{N}(\widehat{\mathbf{x}}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})\mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k,(\beta_k\boldsymbol{\Lambda}_k)^{-1}\right)\,\mathrm{d}\boldsymbol{\mu}_k = \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k,\left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right).$$

Thus we have

$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K}\frac{\alpha_k}{\widehat{\alpha}}\int \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k,\left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k,\nu_k)\,\mathrm{d}\boldsymbol{\Lambda}_k.$$

The final integration over $\boldsymbol{\Lambda}_k$ is the convolution of a Wishart with a Gaussian. Omitting multiplicative constants which are independent of $\widehat{\mathbf{x}}$ we have

$$\begin{aligned}\int \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k, \left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right)\, \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)\,\mathrm{d}\boldsymbol{\Lambda}_k \\ \propto \int |\boldsymbol{\Lambda}_k|^{1/2+(\nu_k-D-1)/2} \exp\left\{-\frac{1}{2\left(1+\beta_k^{-1}\right)}\mathrm{Tr}\left[\boldsymbol{\Lambda}_k(\widehat{\mathbf{x}}-\mathbf{m}_k)(\widehat{\mathbf{x}}-\mathbf{m}_k)^{\mathrm{T}}\right]\right. \\ \left. -\frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Lambda}_k\mathbf{W}_k^{-1}\right]\right\}\mathrm{d}\boldsymbol{\Lambda}_k.\end{aligned}$$

We can now perform this integration by observing that the argument of the integral is an un-normalized Wishart distribution (unsurprisingly since the Wishart is the conjugate prior for the precision of a Gaussian) and so we can write down the result of this integration, up to an overall constant, by using the known normalization coefficient for the Wishart, given by (B.79). Thus we have

$$\begin{aligned}\int \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k, \left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right)\, \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)\,\mathrm{d}\boldsymbol{\Lambda}_k \\ \propto \left|\mathbf{W}_k^{-1}+\frac{1}{\left(1+\beta_k^{-1}\right)}(\widehat{\mathbf{x}}-\mathbf{m}_k)(\widehat{\mathbf{x}}-\mathbf{m}_k)^{\mathrm{T}}\right|^{-(\nu_k+1)/2} \\ \propto \left|\mathbf{I}+\frac{1}{\left(1+\beta_k^{-1}\right)}\mathbf{W}_k(\widehat{\mathbf{x}}-\mathbf{m}_k)(\widehat{\mathbf{x}}-\mathbf{m}_k)^{\mathrm{T}}\right|^{-(\nu_k+1)/2}\end{aligned}$$

where we have omitted factors independent of $\widehat{\mathbf{x}}$ since we are only interested in the functional dependence on $\widehat{\mathbf{x}}$, and we have made use of the result $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$ and omitted an overall factor of $|\mathbf{W}_k^{-1}|$. Next we use the identity

$$\left|\mathbf{I}+\mathbf{a}\mathbf{b}^{\mathrm{T}}\right| = (1+\mathbf{a}^{\mathrm{T}}\mathbf{b})$$

where $\mathbf{a}$ and $\mathbf{b}$ are D-dimensional vectors and $\mathbf{I}$ is the $D \times D$ unit matrix, to give

$$\begin{aligned}\int \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k, \left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right)\, \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)\,\mathrm{d}\boldsymbol{\Lambda}_k \\ \propto \left\{1+\frac{1}{\left(1+\beta_k^{-1}\right)}(\widehat{\mathbf{x}}-\mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\widehat{\mathbf{x}}-\mathbf{m}_k)\right\}^{-(\nu_k+1)/2}.\end{aligned}$$

We recognize this result as being a Student distribution, and by comparison with the standard form (B.68) for the Student we see that it has mean $\mathbf{m}_k$, precision given by (10.82) and degrees of freedom parameter $\nu_k + 1 - D$. We can re-instate the normalization coefficient using the standard form for the Student distribution given in Appendix B.

10.20 Consider first the posterior distribution over the precision of component k given by

$$q^\star(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k).$$

From (10.63) we see that for large N we have $\nu_k \to N_k$, and similarly from (10.62) we see that $\mathbf{W}_k \to N_k^{-1}\mathbf{S}_k^{-1}$. Thus the mean of the distribution over $\boldsymbol{\Lambda}_k$, given by $\mathrm{E}[\boldsymbol{\Lambda}_k] = \nu_k\mathbf{W}_k \to \mathbf{S}_k^{-1}$ which is the maximum likelihood value (this assumes that the quantities r_{nk} reduce to the corresponding EM values, which is indeed the case as we shall show shortly). In order to show that this posterior is also sharply peaked, we consider the differential entropy, $\mathrm{H}[\boldsymbol{\Lambda}_k]$ given by (B.82), and show that, as $N_k \to \infty$, $\mathrm{H}[\boldsymbol{\Lambda}_k] \to 0$, corresponding to the density collapsing to a spike. First consider the normalizing constant $B(\mathbf{W}_k, \nu_k)$ given by (B.79). Since $\mathbf{W}_k \to N_k^{-1}\mathbf{S}_k^{-1}$ and $\nu_k \to N_k$,

$$-\ln B(\mathbf{W}_k, \nu_k) \to -\frac{N_k}{2}\left(D\ln N_k + \ln|\mathbf{S}_k| - D\ln 2\right) + \sum_{i=1}^{D}\ln\Gamma\left(\frac{N_k+1-i}{2}\right).$$

We then make use of Stirling's approximation (1.146) to obtain

$$\ln\Gamma\left(\frac{N_k+1-i}{2}\right) \simeq \frac{N_k}{2}\left(\ln N_k - \ln 2 - 1\right)$$

which leads to the approximate limit

$$\begin{aligned} -\ln B(\mathbf{W}_k, \nu_k) &\to -\frac{N_k D}{2}\left(\ln N_k - \ln 2 - \ln N_k + \ln 2 + 1\right) - \frac{N_k}{2}\ln|\mathbf{S}_k| \\ &= -\frac{N_k}{2}\left(\ln|\mathbf{S}_k| + D\right). \end{aligned} \tag{280}$$

Next, we use (10.241) and (B.81) in combination with $\mathbf{W}_k \to N_k^{-1}\mathbf{S}_k^{-1}$ and $\nu_k \to N_k$ to obtain the limit

$$\begin{aligned} \mathbb{E}\left[\ln|\boldsymbol{\Lambda}|\right] &\to D\ln\frac{N_k}{2} + D\ln 2 - D\ln N_k - \ln|\mathbf{S}_k| \\ &= -\ln|\mathbf{S}_k|, \end{aligned}$$

where we approximated the argument to the digamma function by $N_k/2$. Substituting this and (280) into (B.82), we get

$$\mathrm{H}[\boldsymbol{\Lambda}] \to 0$$

when $N_k \to \infty$.

Next consider the posterior distribution over the mean $\boldsymbol{\mu}_k$ of the k^{th} component given by

$$q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \beta_k\boldsymbol{\Lambda}_k).$$

From (10.61) we see that for large N the mean $\mathbf{m}_k$ of this distribution reduces to $\overline{\mathbf{x}}_k$ which is the corresponding maximum likelihood value. From (10.60) we see that

$\beta_k \to N_k$ and Thus the precision $\beta_k \mathbf{\Lambda}_k \to \beta_k \nu_k \mathbf{W}_k \to N_k \mathbf{S}_k^{-1}$ which is large for large N and hence this distribution is sharply peaked around its mean.

Now consider the posterior distribution $q^\star(\boldsymbol{\pi})$ given by (10.57). For large N we have $\alpha_k \to N_k$ and so from (B.17) and (B.19) we see that the posterior distribution becomes sharply peaked around its mean $\mathrm{E}[\pi_k] = \alpha_k/\overline{\alpha} \to N_k/N$ which is the maximum likelihood solution.

For the distribution $q^\star(\mathbf{z})$ we consider the responsibilities given by (10.67). Using (10.65) and (10.66), together with the asymptotic result for the digamma function, we again obtain the maximum likelihood expression for the responsibilities for large N.

Finally, for the predictive distribution we first perform the integration over $\boldsymbol{\pi}$, as in the solution to Exercise 10.19, to give

$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K} \frac{\alpha_k}{\overline{\alpha}} \iint \mathcal{N}(\widehat{\mathbf{x}}|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) q(\boldsymbol{\mu}_k, \mathbf{\Lambda}_k)\, \mathrm{d}\boldsymbol{\mu}_k\, \mathrm{d}\mathbf{\Lambda}_k.$$

The integrations over $\boldsymbol{\mu}_k$ and $\mathbf{\Lambda}_k$ are then trivial for large N since these are sharply peaked and hence approximate delta functions. We therefore obtain

$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K} \frac{N_k}{N} \mathcal{N}(\widehat{\mathbf{x}}|\overline{\mathbf{x}}_k, \mathbf{W}_k)$$

which is a mixture of Gaussians, with mixing coefficients given by N_k/N.

10.21 The number of equivalent parameter settings equals the number of possible assignments of K parameter sets to K mixture components: K for the first component, times $K-1$ for the second component, times $K-2$ for the third and so on, giving the result $K!$.

10.22 The mixture distribution over the parameter space takes the form

$$q(\mathbf{\Theta}) = \frac{1}{K!} \sum_{\kappa=1}^{K!} q_\kappa(\boldsymbol{\theta}_\kappa)$$

where $\boldsymbol{\theta}_\kappa = \{\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \boldsymbol{\pi}\}$, κ indexes the components of this mixture and $\mathbf{\Theta} = \{\boldsymbol{\theta}_\kappa\}$. With this model, (10.3) becomes

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{\Theta}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{\Theta})}{q(\mathbf{\Theta})} \right\} \mathrm{d}\mathbf{\Theta} \\
&= \frac{1}{K!} \sum_{\kappa=1}^{K!} \int q_\kappa(\boldsymbol{\theta}_\kappa) \ln p(\mathbf{X}, \boldsymbol{\theta}_\kappa)\, \mathrm{d}\boldsymbol{\theta}_\kappa \\
&\qquad - \frac{1}{K!} \sum_{\kappa=1}^{K!} \int q_\kappa(\boldsymbol{\theta}_\kappa) \ln \left(\frac{1}{K!} \sum_{\kappa'=1}^{K!} q_{\kappa'}(\boldsymbol{\theta}_{\kappa'}) \right) \mathrm{d}\boldsymbol{\theta}_\kappa \\
&= \int q(\boldsymbol{\theta}) \ln p(\mathbf{X}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} + \ln K!
\end{aligned}$$

where $q(\boldsymbol{\theta})$ corresponds to any one of the $K!$ equivalent $q_\kappa(\boldsymbol{\theta}_\kappa)$ distributions. Note that in the last step, we use the assumption that the overlap between these distributions is negligible and hence

$$\int q_\kappa(\boldsymbol{\theta}) \ln q_{\kappa'}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \simeq 0$$

when $\kappa \neq \kappa'$.

10.23 When we are treating $\boldsymbol{\pi}$ as a parameter, there is neither a prior, nor a variational posterior distribution, over $\boldsymbol{\pi}$. Therefore, the only term remaining from the lower bound, (10.70), that involves $\boldsymbol{\pi}$ is the second term, (10.72). Note however, that (10.72) involves the *expectations* of $\ln \pi_k$ under $q(\boldsymbol{\pi})$, whereas here, we operate directly with π_k, yielding

$$\mathbb{E}_{q(\mathbf{Z})}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln \pi_k.$$

Adding a Langrange term, as in (9.20), taking the derivative w.r.t. π_k and setting the result to zero we get

$$\frac{N_k}{\pi_k} + \lambda = 0, \tag{281}$$

where we have used (10.51). By re-arranging this to

$$N_k = -\lambda \pi_k$$

and summing both sides over k, we see that $-\lambda = \sum_k N_k = N$, which we can use to eliminate λ from (281) to get (10.83).

10.24 The singularities that may arise in maximum likelihood estimation are caused by a mixture component, k, collapsing on a data point, $\mathbf{x}_n$, i.e., $r_{kn} = 1$, $\boldsymbol{\mu}_k = \mathbf{x}_n$ and $|\boldsymbol{\Lambda}_k| \to \infty$.

However, the prior distribution $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ defined in (10.40) will prevent this from happening, also in the case of MAP estimation. Consider the product of the expected complete log-likelihood and $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ as a function of $\boldsymbol{\Lambda}_k$:

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{Z})}\left[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right] \\
&= \frac{1}{2} \sum_{n=1}^{N} r_{kn} \left(\ln |\boldsymbol{\Lambda}_k| - (\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Lambda}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) \\
&\quad + \ln |\boldsymbol{\Lambda}_k| - \beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0) \\
&\quad + (\nu_0 - D - 1) \ln |\boldsymbol{\Lambda}_k| - \mathrm{Tr}\left[\mathbf{W}_0^{-1} \boldsymbol{\Lambda}_k\right] + \text{const.}
\end{aligned}$$

where we have used (10.38), (10.40) and (10.50), together with the definitions for the Gaussian and Wishart distributions; the last term summarizes terms independent of $\boldsymbol{\Lambda}_k$. Using (10.51)–(10.53), we can rewrite this as

$$(\nu_0 + N_k - D) \ln |\boldsymbol{\Lambda}_k| - \mathrm{Tr}\left[(\mathbf{W}_0^{-1} + \beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + N_k \mathbf{S}_k) \boldsymbol{\Lambda}_k\right],$$

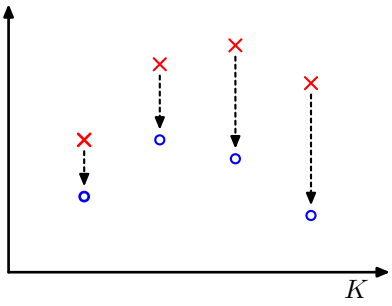


Figure 9 Illustration of the true log marginal likelihood for a Gaussian mixture model (×) and the corresponding variational bound obtained from a factorized approximation (○) as functions of the number of mixture components, K. The dashed arrows emphasize the typical increase in the difference between the true log marginal likelihood and the bound. As a consequence, the bound tends to have its peak at a lower value of K than the true log marginal likelihood.

where we have dropped the constant term. Using (C.24) and (C.28), we can compute the derivative of this w.r.t. $\mathbf{\Lambda}_k$ and setting the result equal to zero, we find the MAP estimate for $\mathbf{\Lambda}_k$ to be

$$\mathbf{\Lambda}_k^{-1} = \frac{1}{\nu_0 + N_k - D}(\mathbf{W}_0^{-1} + \beta_0(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + N_k\mathbf{S}_k).$$

From this we see that $|\mathbf{\Lambda}_k^{-1}|$ can never become 0, because of the presence of $\mathbf{W}_0^{-1}$ (which we must chose to be positive definite) in the expression on the r.h.s.

10.25 As the number of mixture components grows, so does the number of variables that may be correlated, but which are treated as independent under a variational approximation, as illustrated in Figure 10.2. As a result of this, the proportion of probability mass under the true distribution, $p(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}|\mathbf{X})$, that the variational approximation $q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma})$ does not capture, will grow. The consequence will be that the second term in (10.2), the KL divergence between $q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma})$ and $p(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}|\mathbf{X})$, will increase. Since this KL divergence is the difference between the true log marginal and the corresponding the lower bound, the latter must decrease compared to the former. Thus, as illustrated in Figure 9, chosing the number of components based on the lower bound will tend to underestimate the optimal number of components.

10.26 Extending the variational treatment of Section 10.3 to also include β, we specify the prior for β

$$p(\beta) = \mathrm{Gam}\,(\beta|c_0, d_0) \tag{282}$$

and modify (10.90) as

$$p(\mathbf{t}, \mathbf{w}, \alpha, \beta) = p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)p(\alpha)p(\beta) \tag{283}$$

where the first factor on the r.h.s. correspond to (10.87) with the dependence on β made explicit.

The formulae for $q^\star(\alpha)$, (10.93)–(10.95), remain unchanged. For $q(\mathbf{w})$, we follow the path mapped out in Section 10.3, incorporating the modifications required by the

changed treatment of β; (10.96)–(10.98) now become

$$\begin{aligned}\ln q^\star(\mathbf{w}) &= \mathbb{E}_\beta\left[\ln p(\mathbf{t}|\mathbf{w},\beta)\right] + \mathbb{E}_\alpha\left[\ln p(\mathbf{w}|\alpha)\right] + \text{const} \\ &= -\frac{\mathbb{E}[\beta]}{2}\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\phi_n - t_n\right\}^2 - \frac{\mathbb{E}[\alpha]}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const} \\ &= -\frac{1}{2}\mathbf{w}^{\mathrm{T}}\left(\mathbb{E}[\alpha]\mathbf{I} + \mathbb{E}[\beta]\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)\mathbf{w} + \mathbb{E}[\beta]\mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} + \text{const}.\end{aligned}$$

Accordingly, (10.100) and (10.101) become

$$\begin{aligned}\mathbf{m}_N &= \mathbb{E}[\beta]\mathbf{S}_N\mathbf{\Phi}\mathbf{t} \\ \mathbf{S}_N &= \left(\mathbb{E}[\alpha]\mathbf{I} + \mathbb{E}[\beta]\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}.\end{aligned}$$

For $q(\beta)$, we use (10.9), (282) and (283) to obtain

$$\begin{aligned}\ln q^\star(\beta) &= \mathbb{E}_{\mathbf{w}}\left[\ln p(\mathbf{t}|\mathbf{w},\beta)\right] + \ln p(\beta) + \text{const} \\ &= \frac{N}{2}\ln\beta - \frac{\beta}{2}\mathbb{E}_{\mathbf{w}}\left[\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\phi_n - t_n\right\}^2\right] + (c_0 - 1)\ln\beta - d_0\beta\end{aligned}$$

which we recognize as the logarithm of a Gamma distribution with parameters

$$\begin{aligned}c_N &= c_0 + \frac{N}{2} \\ d_N &= d_0 + \frac{1}{2}\mathbb{E}\left[\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\phi_n - t_n\right)^2\right] \\ &= d_0 + \frac{1}{2}\left(\mathrm{Tr}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\right) + \mathbf{t}^{\mathrm{T}}\mathbf{t}\right) - \mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbb{E}[\mathbf{w}] \\ &= d_0 + \frac{1}{2}\left(\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2 + \mathrm{Tr}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{S}_N\right)\right)\end{aligned}$$

where we have used (10.103) and, from (B.38),

$$\mathbb{E}[\mathbf{w}] = \mathbf{m}_N. \tag{284}$$

Thus, from (B.27),

$$\mathbb{E}[\beta] = \frac{c_N}{d_N}. \tag{285}$$

In the lower bound, (10.107), the first term will be modified and two new terms added on the r.h.s.We start with the modified log likelihood term:

$$\begin{aligned}\mathbb{E}_\beta\left[\mathbb{E}_{\mathbf{w}}\left[\ln p(\mathbf{t}|\mathbf{w},\beta)\right]\right] &= \frac{N}{2}\left(\mathbb{E}[\beta] - \ln(2\pi)\right) - \frac{\mathbb{E}[\beta]}{2}\mathbb{E}\left[\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2\right] \\ &= \frac{N}{2}\left(\psi(c_N) - \ln d_N - \ln(2\pi)\right) \\ &\quad -\frac{c_N}{2d_N}\left(|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 + \mathrm{Tr}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{S}_N\right)\right)\end{aligned}$$

where we have used (284), (285), (10.103) and (B.30). Next we consider the term corresponding log prior over β:

$$\begin{aligned}\mathbb{E}\left[\ln p(\beta)\right] &= (c_0-1)\mathbb{E}[\ln\beta] - d_0\mathbb{E}[\beta] + c_0\ln d_0 - \ln\Gamma(c_0) \\ &= (c_0-1)(\psi(c_N)-\ln d_N) - \frac{d_0 c_N}{d_N} + c_0\ln d_0 - \ln\Gamma(c_0)\end{aligned}$$

where we have used (285) and (B.30). Finally, from (B.31), we get the last term in the form of the negative entropy of the posterior over β:

$$-\mathbb{E}\left[\ln q^\star(\beta)\right] = (c_N-1)\psi(c_N) + \ln d_N - c_N - \ln\Gamma(c_N).$$

Finally, the predictive distribution is given by (10.105) and (10.106), with $1/\beta$ replaced by $1/\mathbb{E}[\beta]$.

10.27 Consider each of the five terms in the lower bound (10.107) in turn. For the terms arising from the likelihood function we have

$$\begin{aligned}\mathbb{E}[\ln p(\mathbf{t}|\mathbf{w})] &= -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{\beta}{2}\mathbb{E}\left[\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)^2\right] \\ &= -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta \\ &\quad -\frac{\beta}{2}\left\{\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbb{E}[\mathbf{w}^{\mathrm{T}}]\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \mathrm{Tr}\left(\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\right\}.\end{aligned}$$

The prior over $\mathbf{w}$ gives rise to

$$\mathbb{E}[\ln p(\mathbf{w}|\alpha)] = -\frac{M}{2}\ln(2\pi) + \frac{M}{2}\mathbb{E}[\ln\alpha] - \frac{\mathbb{E}[\alpha]}{2}\mathbb{E}[\mathbf{w}^{\mathrm{T}}\mathbf{w}].$$

Similarly, the prior over α gives

$$\mathbb{E}[\ln p(\alpha)] = a_0\ln b_0 + (a_0-1)\mathbb{E}[\ln\alpha] - b_0\mathbb{E}[\alpha] - \ln\Gamma(a_0).$$

The final two terms in $\mathcal{L}$ represent the negative entropies of the Gaussian and gamma distributions, and are given by (B.41) and (B.31) respectively, so that

$$-\mathbb{E}[\ln q(\mathbf{w})] = \frac{1}{2}\ln|\mathbf{S}_N| + \frac{M}{2}(1+\ln(2\pi)).$$

Similarly we have

$$-\mathbb{E}[\ln q(\alpha)] = -(a_n-1)\psi(a_N) + a_N + \ln\Gamma(a_N) + \ln b_N.$$

Now we substitute the following expressions for the moments

$$\begin{aligned}\mathbb{E}[\mathbf{w}] &= \mathbf{m}_N \\ \mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] &= \mathbf{m}_N\mathbf{m}_N^{\mathrm{T}} + \mathbf{S}_N \\ \mathbb{E}[\alpha] &= \frac{a_N}{b_N} \\ \mathbb{E}[\ln\alpha] &= \psi(a_N) - \ln b_N.\end{aligned}$$

and combine the various terms together to obtain (10.107).

10.28 **NOTE**: In PRML, Equations (10.119)–(10.121) contain errors; please consult the PRML Errata for relevant corrections.

We start by writing the complete-data likelihood, given by (10.37) and (10.38) in a form corresponding to (10.113). From (10.37) and (10.38), we have

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\mathbf{Z}|\boldsymbol{\pi}) \\ &= \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right)^{z_{nk}} \end{aligned}$$

which is a product over data points, just like (10.113). Focussing on the individual factors of this product, we have

$$p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right)^{z_{nk}} = \exp\left\{\sum_{k=1}^{K} z_{nk} \left(\ln \pi_k + \frac{1}{2} \ln |\boldsymbol{\Lambda}_k| - \frac{D}{2} \ln(2\pi) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right)\right\}.$$

Drawing on results from Solution 2.57, we can rewrite this in the form of (10.113), with

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k \\ \overrightarrow{\boldsymbol{\Lambda}_k} \\ \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k \\ \ln |\boldsymbol{\Lambda}_k| \\ \ln \pi_k \end{bmatrix}_{k=1,\ldots,K} \tag{286}$$

$$\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) = \left[z_{nk} \begin{bmatrix} \mathbf{x}_n \\ \overrightarrow{\frac{1}{2}\mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}} \\ -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{bmatrix} \right]_{k=1,\ldots,K} \tag{287}$$

$$h(\mathbf{x}_n, \mathbf{z}_n) = \prod_{k=1}^{K} \left((2\pi)^{-D/2}\right)^{z_{nk}} \tag{288}$$

$$g(\boldsymbol{\eta}) = 1$$

where we have introduce the notation

$$[\mathbf{v}_k]_{k=1,\ldots,K} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_K \end{bmatrix}$$

and the operator $\overrightarrow{\mathbf{M}}$ which returns a vector formed by stacking the columns of the argument matrix on top of each other.

Next we seek to rewrite the prior over the parameters, given by (10.39) and (10.40), in a form corresponding to (10.114) and which also matches (286). From, (10.39), (10.40) and Appendix B, we have

$$\begin{aligned} p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \mathrm{Dir}\left(\boldsymbol{\pi}|\boldsymbol{\alpha}_0\right) \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_0, \left(\beta_0\boldsymbol{\Lambda}_k\right)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0\right) \\ &= C(\boldsymbol{\alpha}_0) \left(\frac{\beta_0}{2\pi}\right)^{KD/2} B(\mathbf{W}_0, \nu_0)^K \exp\left\{ \sum_{k=1}^{K} (\alpha_0 - 1) \ln \pi_k \right. \\ &\left. + \frac{\nu_0 - D}{2} \ln |\boldsymbol{\Lambda}_k| - \frac{1}{2} \mathrm{Tr}\left(\boldsymbol{\Lambda}_k \left[\beta_0(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + \mathbf{W}_0\right]\right) \right\}. \end{aligned}$$

we can rewrite this in the form of (10.114) with $\boldsymbol{\eta}$ given by (286),

$$\boldsymbol{\chi}_0 = \begin{bmatrix} \beta_0 \mathbf{m}_0 \\ -\frac{1}{2}\left(\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}}\right) \\ -\beta_0/2 \\ (\nu_0 - D)/2 \\ \alpha_0 - 1 \end{bmatrix}_{k=1,\ldots,K} \tag{289}$$

$$g(\boldsymbol{\eta}) = 1$$

$$f(\upsilon_0, \boldsymbol{\chi}_0) = C(\boldsymbol{\alpha}_0) \left(\frac{\beta_0}{2\pi}\right)^{KD/2} B(\mathbf{W}_0, \nu_0)^K$$

and υ_0 replaces ν_0 in (10.114) to avoid confusion with ν_0 in (10.40).

Having rewritten the Bayesian mixture of Gaussians as a conjugate model from the exponential family, we now proceed to rederive (10.48), (10.57) and (10.59). By exponentiating both sides of (10.115) and making use of (286)–(288), we obtain (10.47), with ρ_{nk} given by (10.46), from which (10.48) follows.

Next we can use (10.50) to take the expectation w.r.t. $\mathbf{Z}$ in (10.121), substituting r_{nk} for $\mathbb{E}[z_{nk}]$ in (287). Combining this with (289), (10.120) and (10.121) become

$$\upsilon_N = \upsilon_0 + N = 1 + N$$

and

$$\upsilon_N \boldsymbol{\chi}_N = \left[\begin{array}{c} \beta_0 \mathbf{m}_0 \\ -\frac{1}{2}\left(\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}}\right) \\ -\beta_0/2 \\ (\nu_0 - D)/2 \\ \alpha_0 - 1 \end{array}\right]_{k=1,\ldots,K} + \sum_{n=1}^{N} \left[r_{nk} \left[\begin{array}{c} \mathbf{x}_n \\ \frac{1}{2}\overrightarrow{\mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}} \\ -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{array}\right]\right]_{k=1,\ldots,K}$$

$$= \left[\begin{array}{c} \beta_0 \mathbf{m}_0 + N_k \overline{\mathbf{x}}_k \\ -\frac{1}{2}\left(\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}} + N_k \overrightarrow{\left(\mathbf{S}_k + \overline{\mathbf{x}}_k \overline{\mathbf{x}}_k^{\mathrm{T}}\right)}\right) \\ -(\beta_0 + N_k)/2 \\ (\nu_0 - D + N_k)/2 \\ \alpha_0 - 1 + N_k \end{array}\right]_{k=1,\ldots,K} \tag{290}$$

where we use υ_N instead of ν_N in (10.119)–(10.121), to avoid confusion with ν_k, which appears in (10.59) and (10.63). From the bottom row of (287) and (290), we see that the inner product of $\boldsymbol{\eta}$ and $\upsilon_N \boldsymbol{\chi}_N$ gives us the r.h.s. of (10.56), from which (10.57) follows. The remaining terms of this inner product are

$$\sum_{k=1}^{K} \Big\{ \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \left(\beta_0 \mathbf{m}_0 + N_k \overline{\mathbf{x}}_k\right) - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k \left[\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}} + N_k \overrightarrow{\left(\mathbf{S}_k + \overline{\mathbf{x}}_k \overline{\mathbf{x}}_k^{\mathrm{T}}\right)}\right]\right) - \frac{1}{2}(\beta_0 + N_k)\boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k + \frac{1}{2}(\nu_0 + N_k - D) \ln |\boldsymbol{\Lambda}| \Big\}.$$

Restricting our attention to parameters corresponding to a single mixture component and making use of (10.60), (10.61) and (10.63), we can rewrite this as

$$-\frac{1}{2}\beta_k \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k + \beta_k \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \mathbf{m}_k - \frac{1}{2}\beta_k \mathbf{m}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \mathbf{m}_k + \frac{1}{2}\ln |\boldsymbol{\Lambda}| + \frac{1}{2}\beta_k \mathbf{m}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \mathbf{m}_k - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k \left[\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}} + N_k \overrightarrow{\left(\mathbf{S}_k + \overline{\mathbf{x}}_k \overline{\mathbf{x}}_k^{\mathrm{T}}\right)}\right]\right) + \frac{1}{2}(\nu_k - D - 1) \ln |\boldsymbol{\Lambda}|.$$

The first four terms match the logarithm of $\mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}\right)$ from the r.h.s. of (10.59); the missing $D/2[\ln \beta_k - \ln(2\pi)]$ can be accounted for in $f(\upsilon_N, \boldsymbol{\chi}_N)$. To make the remaining terms match the logarithm of $\mathcal{W}\left(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0\right)$ from the r.h.s. of (10.59), we need to show that

$$\beta_0 \mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}} + N_k \overline{\mathbf{x}}_k \overline{\mathbf{x}}_k^{\mathrm{T}} - \beta_k \mathbf{m}_k \mathbf{m}_k^{\mathrm{T}}$$

equals the last term on the r.h.s. of (10.62). Using (10.60), (10.61) and (276), we get

$$\begin{aligned}
&\beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - \beta_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\
&= \beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - \beta_0\mathbf{m}_0\mathbf{m}_k^{\mathrm{T}} - N_k\overline{\mathbf{x}}_k\mathbf{m}_k^{\mathrm{T}} \\
&= \beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} - \beta_0\mathbf{m}_0\mathbf{m}_k^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - N_k\overline{\mathbf{x}}_k\mathbf{m}_k^{\mathrm{T}} + \beta_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} - \beta_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\
&= \beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} - \beta_0\mathbf{m}_0\mathbf{m}_k^{\mathrm{T}} - \beta_0\mathbf{m}_k\mathbf{m}_0^{\mathrm{T}} + \beta_0\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\
&\quad + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - N_k\overline{\mathbf{x}}_k\mathbf{m}_k^{\mathrm{T}} - N_k\mathbf{m}_k\overline{\mathbf{x}}_k^{\mathrm{T}} + N_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\
&= \beta_0(\mathbf{m}_k - \mathbf{m}_0)(\mathbf{m}_k - \mathbf{m}_0)^{\mathrm{T}} + N_k(\overline{\mathbf{x}}_k - \mathbf{m}_k)(\overline{\mathbf{x}}_k - \mathbf{m}_k)^{\mathrm{T}} \\
&= \frac{\beta_0 N_k}{\beta_0 + N_k}(\overline{\mathbf{x}}_k - \mathbf{m}_0)(\overline{\mathbf{x}}_k - \mathbf{m}_0)^{\mathrm{T}}.
\end{aligned}$$

Thus we have recovered $\ln \mathcal{W}(\mathbf{\Lambda}_k|\mathbf{W}_0, \nu_0)$ (missing terms are again accounted for by $f(\upsilon_N, \boldsymbol{\chi}_N)$) and thereby (10.59).

10.29 **NOTE**: In the 1st printing of PRML, the use of λ to denote the varitional parameter leads to inconsistencies w.r.t. exisiting literature. To remedy this λ should be replaced by η from the beginning of Section 10.5 up to and including the last line before equation (10.141). For further details, please consult the PRML Errata.

Standard rules of differentiation give

$$\frac{d\ln(x)}{dx} = \frac{1}{x}$$

$$\frac{d^2\ln(x)}{dx^2} = -\frac{1}{x^2}.$$

Since its second derivative is negative for all value of x, $\ln(x)$ is concave for $0 < x < \infty$.

From (10.133) we have

$$\begin{aligned}
g(\eta) &= \min_x \{\eta x - f(x)\} \\
&= \min_x \{\eta x - \ln(x)\}.
\end{aligned}$$

We can minimize this w.r.t. x by setting the corresponding derivative to zero and solving for x:

$$\frac{dg}{dx} = \eta - \frac{1}{x} = 0 \quad \Longrightarrow \quad x = \frac{1}{\eta}.$$

Substituting this in (10.133), we see that

$$g(\eta) = 1 - \ln\left(\frac{1}{\eta}\right).$$

If we substitute this into (10.132), we get

$$f(x) = \min_\eta \left\{\eta x - 1 + \ln\left(\frac{1}{\eta}\right)\right\}.$$

Again, we can minimize this w.r.t. η by setting the corresponding derivative to zero and solving for η:

$$\frac{df}{d\eta} = x - \frac{1}{\eta} = 0 \quad \Longrightarrow \quad \eta = \frac{1}{x},$$

and substituting this into (10.132), we find that

$$f(x) = \frac{1}{x}x - 1 + \ln\left(\frac{1}{1/x}\right) = \ln(x).$$

10.30 **NOTE**: Please consult note preceding Solution 10.29 for relevant corrections.

Differentiating the log logistic function, we get

$$\frac{d}{dx}\ln\sigma = \left(1 + e^{-x}\right)^{-1} e^{-x} = \sigma(x)e^{-x} \tag{291}$$

and, using (4.88),

$$\frac{d^2}{dx^2}\ln\sigma = \sigma(x)(1-\sigma(x))e^{-x} - \sigma(x)e^{-x} = -\sigma(x)^2 e^{-x}$$

which will always be negative and hence $\ln\sigma(x)$ is concave.

From (291), we see that the first order Taylor expansion of $\ln\sigma(x)$ around ξ becomes

$$\ln\sigma(x) = \ln\sigma(\xi) + (x-\xi)\sigma(\xi)e^{-\xi} + O\left((x-xi)^2\right).$$

Since $\ln\sigma(x)$ is concave, its tangent line will be an upper bound and hence

$$\ln\sigma(x) \leqslant \ln\sigma(\xi) + (x-\xi)\sigma(\xi)e^{-\xi}. \tag{292}$$

Following the presentation in Section 10.5, we define

$$\eta = \sigma(\xi)e^{-\xi}. \tag{293}$$

Using (4.60), we have

$$\begin{aligned}
\eta &= \sigma(\xi)e^{-\xi} = \frac{e^{-\xi}}{1+e^{-\xi}} \\
&= \frac{1}{1+e^{\xi}} = \sigma(-\xi) \\
&= 1 - \sigma(\xi)
\end{aligned}$$

and hence

$$\sigma(\xi) = 1 - \eta.$$

From this and (293)

$$\begin{aligned}
\xi &= \ln\sigma(\xi) - \ln\eta \\
&= \ln(1-\eta) - \ln\eta.
\end{aligned}$$

Using these results in (292), we have

$$\ln \sigma(x) \leqslant \ln(1-\eta) + x\eta - \eta\left[\ln(1-\eta) - \ln \eta\right].$$

By exponentiating both sides and making use of (10.135), we obtain (10.137).

10.31 **NOTE**: Please consult note preceding Solution 10.29 for relevant corrections.

Taking the derivative of $f(x)$ w.r.t. x we get

$$\frac{\mathrm{d}f}{\mathrm{d}x} = -\frac{1}{e^{x/2}+e^{-x/2}}\frac{1}{2}\left(e^{x/2}-e^{-x/2}\right) = -\frac{1}{2}\tanh\left(\frac{x}{2}\right)$$

where we have used (5.59). From (5.60), we get

$$f''(x) = \frac{d^2 f}{dx^2} = -\frac{1}{4}\left(1-\tanh\left(\frac{x}{2}\right)^2\right).$$

Since $\tanh(x/2)^2 < 1$ for finite values of x, $f''(x)$ will always be negative and so $f(x)$ is concave.

Next we define $y = x^2$, noting that y will always be non-negative, and express f as a function of y:

$$f(y) = -\ln\left\{\exp\left(\frac{\sqrt{y}}{2}\right) + \exp\left(-\frac{\sqrt{y}}{2}\right)\right\}.$$

We then differentiate f w.r.t. y, yielding

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}y} &= -\left\{\exp\left(\frac{\sqrt{y}}{2}\right) + \exp\left(-\frac{\sqrt{y}}{2}\right)\right\}^{-1} \\ &\qquad \frac{1}{4\sqrt{y}}\left\{\exp\left(\frac{\sqrt{y}}{2}\right) - \exp\left(-\frac{\sqrt{y}}{2}\right)\right\} \qquad (294)\\ &= -\frac{1}{4\sqrt{y}}\tanh\left(\frac{\sqrt{y}}{2}\right). \qquad (295)\end{aligned}$$

and, using (5.60),

$$\begin{aligned}\frac{\mathrm{d}^2 f}{\mathrm{d}y^2} &= \frac{1}{8y^{3/2}}\tanh\left(\frac{\sqrt{y}}{2}\right) - \frac{1}{16y}\left\{1-\tanh\left(\frac{\sqrt{y}}{2}\right)^2\right\} \\ &= \frac{1}{8y}\left(\tanh\left(\frac{\sqrt{y}}{2}\right)\left\{\frac{1}{\sqrt{y}} + \frac{1}{2}\tanh\left(\frac{\sqrt{y}}{2}\right)\right\} - \frac{1}{2}\right). \qquad (296)\end{aligned}$$

We see that this will be positive if the factor inside the outermost parenthesis is positive, which is equivalent to

$$\frac{1}{\sqrt{y}}\tanh\left(\frac{\sqrt{y}}{2}\right) > \frac{1}{2}\left\{1-\tanh^2\left(\frac{\sqrt{y}}{2}\right)\right\}.$$

If we divide both sides by $\tanh\left(\sqrt{y}/2\right)$, substitute a for $\sqrt{y}/2$ and then make use of (5.59), we can write this as

$$\begin{aligned}\frac{1}{a} &> \frac{e^a + e^{-a}}{e^a - e^{-a}} - \frac{e^a - e^{-a}}{e^a + e^{-a}} \\ &= \frac{\left(e^a + e^{-a}\right)^2 - \left(e^a - e^{-a}\right)^2}{\left(e^a - e^{-a}\right)\left(e^a + e^{-a}\right)} \\ &= \frac{4}{e^{2a} - e^{-2a}}.\end{aligned}$$

Taking the inverse of both sides of this inequality we get

$$a < \frac{1}{4}\left(e^{2a} - e^{-2a}\right).$$

If differentiate both sides w.r.t. a we see that the derivatives are equal at $a = 0$ and for $a > 0$, the derivative of the r.h.s. will be greater than that of the l.h.s. Thus, the r.h.s. will grow faster and the inequality will hold for $a > 0$. Consequently (296) will be positive for $y > 0$ and approach $+\infty$ as y approaches 0.

Now we use (295) to make a Taylor expansion of $f(x^2)$ around ξ^2, which gives

$$\begin{aligned}f(x^2) &= f(\xi^2) + (x^2 - \xi^2)f'(\xi^2) + O\left((x^2 - \xi^2)^2\right) \\ &\geqslant -\ln\left\{\exp\left(\frac{\xi}{2}\right) + \exp\left(-\frac{\xi}{2}\right)\right\} - (x^2 - \xi^2)\frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right).\end{aligned}$$

where we have used the fact that f is convex function of x^2 and hence its tangent will be a lower bound. Defining

$$\lambda(\xi) = \frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right)$$

we recover (10.143), from which (10.144) follows.

10.32 We can see this from the lower bound (10.154), which is simply a sum of the prior and indepedent contributions from the data points, all of which are quadratic in $\mathbf{w}$. A new data point would simply add another term to this sum and we can regard terms from the previously arrived data points and the original prior collectively as a revised prior, which should be combined with the contributions from the new data point.

The corresponding sufficient statistics, (10.157) and (10.158), can be rewritten di-

rectly in the corresponding sequential form,

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N}(t_n - 1/2)\boldsymbol{\phi}_n\right) \\
&= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N-1}(t_n - 1/2)\boldsymbol{\phi}_n + (t_N - 1/2)\boldsymbol{\phi}_N\right) \\
&= \mathbf{S}_N\left(\mathbf{S}_{N-1}^{-1}\mathbf{S}_{N-1}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N-1}(t_n - 1/2)\boldsymbol{\phi}_n\right) + (t_N - 1/2)\boldsymbol{\phi}_N\right) \\
&= \mathbf{S}_N\left(\mathbf{S}_{N-1}^{-1}\mathbf{m}_{N-1} + (t_N - 1/2)\boldsymbol{\phi}_N\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}} \\
&= \mathbf{S}_0^{-1} + 2\sum_{n=1}^{N-1}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}} + 2\lambda(\xi_N)\boldsymbol{\phi}_N\boldsymbol{\phi}_N^{\mathrm{T}} \\
&= \mathbf{S}_{N-1}^{-1} + 2\lambda(\xi_N)\boldsymbol{\phi}_N\boldsymbol{\phi}_N^{\mathrm{T}}.
\end{aligned}$$

The update formula for the variational parameters, (10.163), remain the same, but each parameter is updated only once, although this update will be part of an iterative scheme, alternating between updating $\mathbf{m}_N$ and $\mathbf{S}_N$ with ξ_N kept fixed, and updating ξ_N with $\mathbf{m}_N$ and $\mathbf{S}_N$ kept fixed. Note that updating ξ_N will not affect $\mathbf{m}_{N-1}$ and $\mathbf{S}_{N-1}$. Note also that this updating policy differs from that of the batch learning scheme, where all variational parameters are updated using statistics based on all data points.

10.33 Taking the derivative of (10.161) w.r.t. ξ_n, we get

$$\begin{aligned}
\frac{\partial Q}{\partial \xi_n} &= \frac{1}{\sigma(\xi_n)}\sigma'(\xi_n) - \frac{1}{2} - \lambda'(\xi_n)\left(\phi_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\phi - \xi_n^2\right) + \lambda(\xi_n)2\xi_n \\
&= \frac{1}{\sigma(\xi_n)}\sigma(\xi_n)(1-\sigma(x)) - \frac{1}{2} - \lambda'(\xi_n)\left(\phi_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\phi - \xi_n^2\right) \\
&\qquad + \frac{1}{2\xi_n}\left[\sigma(x) - \frac{1}{2}\right]2\xi_n \\
&= -\lambda'(\xi_n)\left(\phi_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\phi - \xi_n^2\right)
\end{aligned}$$

where we have used (4.88) and (10.141). Setting this equal to zero, we obtain (10.162), from which (10.163) follows.

10.34 **NOTE**: In the 1st printing of PRML, there are a number of sign errors in Equation (10.164); the correct form is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\xi}) &= \frac{1}{2}\ln\frac{|\mathbf{S}_N|}{|\mathbf{S}_0|}+\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N-\frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 \\ &\quad+\sum_{n=1}^{N}\left\{\ln\sigma(\xi_n)-\frac{1}{2}\xi_n+\lambda(\xi_n)\xi_n^2\right\}.\end{aligned}$$

We can differentiate $\mathcal{L}$ w.r.t. ξ_n using (3.117) and results from Solution 10.33, to obtain

$$\frac{\partial\mathcal{L}}{\partial\xi_n}=\frac{1}{2}\mathrm{Tr}\left(\mathbf{S}_N^{-1}\frac{\partial\mathbf{S}_N}{\partial\xi_n}\right)+\frac{1}{2}\mathrm{Tr}\left(\mathbf{a}_N\mathbf{a}_N^{\mathrm{T}}\frac{\partial\mathbf{S}_N}{\partial\xi_n}\right)+\lambda'(\xi_n)\xi_n^2 \tag{297}$$

where we have defined

$$\mathbf{a}_N=\mathbf{S}_N^{-1}\mathbf{m}_N. \tag{298}$$

From (10.158) and (C.21), we get

$$\begin{aligned}\frac{\partial\mathbf{S}_N}{\partial\xi_n} &= \frac{\partial\left(\mathbf{S}_N^{-1}\right)^{-1}}{\partial\xi_n}=-\mathbf{S}_N\frac{\partial\mathbf{S}_N^{-1}}{\partial\xi_n}\mathbf{S}_N \\ &= -\mathbf{S}_N 2\lambda'(\xi_n)\phi_n\phi_n^{\mathrm{T}}\mathbf{S}_N.\end{aligned}$$

Substituting this into (297) and setting the result equal to zero, we get

$$-\frac{1}{2}\mathrm{Tr}\left(\left(\mathbf{S}_N^{-1}+\mathbf{a}_N\mathbf{a}_N^{\mathrm{T}}\right)\mathbf{S}_N 2\lambda'(\xi_n)\phi_n\phi_n^{\mathrm{T}}\mathbf{S}_N\right)+\lambda'(\xi_n)\xi_n^2=0.$$

Rearranging this and making use of (298) we get

$$\begin{aligned}\xi_n^2 &= \phi_n^{\mathrm{T}}\mathbf{S}_N\left(\mathbf{S}_N^{-1}+\mathbf{a}_N\mathbf{a}_N^{\mathrm{T}}\right)\mathbf{S}_N\phi_n \\ &= \phi_n^{\mathrm{T}}\left(\mathbf{S}_N+\mathbf{m}_N\mathbf{m}_N^{\mathrm{T}}\right)\phi_n\end{aligned}$$

where we have also used the symmetry of $\mathbf{S}_N$.

10.35 **NOTE**: See note in Solution 10.34.

From (2.43), (4.140) and (10.153), we see that

$$\begin{aligned}p(\mathbf{w})h(\mathbf{w},\boldsymbol{\xi}) &= (2\pi)^{-W/2}|\mathbf{S}_0|^{-1/2} \\ &\quad\exp\left\{-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}+2\sum_{n=1}^{N}\lambda(\xi_n)\phi_n\phi_n^{\mathrm{T}}\right)\mathbf{w}\right. \\ &\qquad\left.+\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}\mathbf{m}_0+\sum_{n=1}^{N}\phi_n\left[t_n-\frac{1}{2}\right]\right)\right\} \\ &\quad\exp\left\{-\frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0+\sum_{n=1}^{N}\frac{\xi_n}{2}+\lambda(\xi_n)\xi_n^2\right\}\prod_{n=1}^{N}\sigma(\xi_n).\end{aligned}$$

Using (10.157) and (10.158), we can complete the square over $\mathbf{w}$, yielding

$$\begin{aligned} p(\mathbf{w})h(\mathbf{w},\boldsymbol{\xi}) &= (2\pi)^{-W/2}\,|\mathbf{S}_0|^{-1/2}\prod_{n=1}^{N}\sigma(\xi_n) \\ &\quad \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right\} \\ &\quad \exp\left\{\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 \sum_{n=1}^{N}\frac{\xi_n}{2} + \lambda(\xi_n)\xi_n^2\right\}. \end{aligned}$$

Now we can do the integral over $\mathbf{w}$ in (10.159), in effect replacing the first exponential factor with $(2\pi)^{W/2}\,|\mathbf{S}_N|^{1/2}$. Taking logarithm, we then obtain (10.164).

10.36 If we denote the joint distribution corresponding to the first j factors by $p_j(\boldsymbol{\theta},\mathcal{D})$, with corresponding evidence $p_j(\mathcal{D})$, then we have

$$\begin{aligned} p_j(\mathcal{D}) &= \int p_j(\boldsymbol{\theta},\mathcal{D})\,\mathrm{d}\boldsymbol{\theta} = \int p_{j-1}(\boldsymbol{\theta},\mathcal{D})f_j(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= p_{j-1}(\mathcal{D})\int p_{j-1}(\boldsymbol{\theta}|\mathcal{D})f_j(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &\simeq p_{j-1}(\mathcal{D})\int q_{j-1}(\boldsymbol{\theta})f_j(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = p_{j-1}(\mathcal{D})Z_j. \end{aligned}$$

By applying this result recursively we see that the evidence is given by the product of the normalization constants

$$p(\mathcal{D}) = \prod_j Z_j.$$

10.37 Here we use the general expectation-propagation equations (10.204)–(10.207). The initial $q(\boldsymbol{\theta})$ takes the form

$$q_{\text{init}}(\boldsymbol{\theta}) = \widetilde{f}_0(\boldsymbol{\theta})\prod_{i\neq 0}\widetilde{f}_i(\boldsymbol{\theta})$$

where $\widetilde{f}_0(\boldsymbol{\theta}) = f_0(\boldsymbol{\theta})$. Thus

$$q^{\backslash 0}(\boldsymbol{\theta}) \propto \prod_{i\neq 0}\widetilde{f}_i(\boldsymbol{\theta})$$

and $q^{\text{new}}(\boldsymbol{\theta})$ is determined by matching moments (sufficient statistics) against

$$q^{\backslash 0}(\boldsymbol{\theta})f_0(\boldsymbol{\theta}) = q_{\text{init}}(\boldsymbol{\theta}).$$

Since by definition this belongs to the same exponential family form as $q^{\text{new}}(\boldsymbol{\theta})$ it follows that

$$q^{\text{new}}(\boldsymbol{\theta}) = q_{\text{init}}(\boldsymbol{\theta}) = q^{\backslash 0}(\boldsymbol{\theta})f_0(\boldsymbol{\theta}).$$

Thus

$$\widetilde{f}_0(\boldsymbol{\theta}) = \frac{Z_0\, q^{\text{new}}(\boldsymbol{\theta})}{q^{\backslash 0}(\boldsymbol{\theta})} = Z_0 f_0(\boldsymbol{\theta})$$

where

$$Z_0 = \int q^{\backslash 0}(\boldsymbol{\theta}) f_0(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = \int q^{\text{new}}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = 1.$$

10.38 The ratio is given by

$$\begin{aligned} q^{\backslash n}(\boldsymbol{\theta}) &\propto \exp\left\{-\frac{1}{2v}\|\boldsymbol{\theta}-\mathbf{m}\|^2 + \frac{1}{2v_n}\|\boldsymbol{\theta}-\mathbf{m}_n\|^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}\left(\frac{1}{v}-\frac{1}{v_n}\right) + \boldsymbol{\theta}^{\mathrm{T}}\left(\frac{1}{v}\mathbf{m}-\frac{1}{v_n}\mathbf{m}_n\right)\right\} \end{aligned}$$

from which we obtain the variance given by (10.215). The mean is then obtained by completing the square and is therefore given by

$$\begin{aligned} \mathbf{m}^{\backslash n} &= v^{\backslash n}\left(v^{-1}\mathbf{m} - v_n^{-1}\mathbf{m}_n\right) \\ &= v^{\backslash n}\left(v^{-1}\mathbf{m} - v_n^{-1}\mathbf{m}_n\right) + v^{\backslash n}v_n^{-1}\mathbf{m} - v^{\backslash n}v_n^{-1}\mathbf{m} \\ &= v^{\backslash n}\left(v^{-1} - v_n^{-1}\right)\mathbf{m} + v^{\backslash n}v_n^{-1}\left(\mathbf{m}-\mathbf{m}_n\right) \end{aligned}$$

Hence we obtain (10.214).

The normalization constant is given by

$$Z_n = \int \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}^{\backslash n}, v^{\backslash n}\mathbf{I})\left\{(1-w)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\theta},\mathbf{I}) + w\mathcal{N}(\mathbf{x}_n|\mathbf{0}, a\mathbf{I})\right\}\,\mathrm{d}\boldsymbol{\theta}.$$

The first term can be integrated by using the result (2.115) while the second term is trivial since the background distribution does not depend on $\boldsymbol{\theta}$ and hence can be taken outside the integration. We therefore obtain (10.216).

10.39 **NOTE**: In PRML, a term $v^{\backslash n}D$ should be added to the r.h.s. of (10.245).

We derive (10.244) by noting

$$\begin{aligned} \nabla_{\mathbf{m}^{\backslash n}} \ln Z_n &= \frac{1}{Z_n}\nabla_{\mathbf{m}^{\backslash n}} \int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{Z_n}\int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\left\{-\frac{1}{v^{\backslash n}}(\mathbf{m}^{\backslash n}-\boldsymbol{\theta})\right\}\,\mathrm{d}\boldsymbol{\theta} \\ &= -\frac{\mathbf{m}^{\backslash n}}{v^{\backslash n}} + \frac{\mathbb{E}[\boldsymbol{\theta}]}{v^{\backslash n}}. \end{aligned}$$

We now use this to derive (10.217) by substituting for Z_n using (10.216) to give

$$\begin{aligned} \nabla_{\mathbf{m}^{\backslash n}} \ln Z_n &= \frac{1}{Z_n}(1-w)\mathcal{N}(\mathbf{x}_n|\mathbf{m}^{\backslash n}, (v^{\backslash n}+1)\mathbf{I})\frac{1}{v^{\backslash n}+1}(\mathbf{x}_n - \mathbf{m}^{\backslash n}) \\ &= \rho_n \frac{1}{v^{\backslash n}+1}(\mathbf{x}_n - \mathbf{m}^{\backslash n}) \end{aligned}$$

where we have defined

$$\rho_n = (1-w)\frac{1}{Z_n}\mathcal{N}(\mathbf{x}_n|\mathbf{m}^{\backslash n},(v^{\backslash n}+1)\mathbf{I}) = 1 - \frac{w}{Z_n}\mathcal{N}(\mathbf{x}_n|\mathbf{0},a\mathbf{I}).$$

Similarly for (10.245) we have

$$\begin{aligned}\nabla_{v^{\backslash n}} \ln Z_n &= \frac{1}{Z_n}\nabla_{v^{\backslash n}} \int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{Z_n}\int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\left\{\frac{1}{2(v^{\backslash n})^2}\left(\mathbf{m}^{\backslash n}-\boldsymbol{\theta}\right)^{\mathrm{T}}\left(\mathbf{m}^{\backslash n}-\boldsymbol{\theta}\right) - \frac{D}{2v^{\backslash n}}\right\}\mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{2(v^{\backslash n})^2}\left\{\mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}] - 2\mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}]\mathbf{m}^{\backslash n} + \|\mathbf{m}^{\backslash n}\|^2\right\} - \frac{D}{2v^{\backslash n}}.\end{aligned}$$

Re-arranging we obtain (10.245). Now we substitute for Z_n using (10.216) to give

$$\begin{aligned}\nabla_{v^{\backslash n}} \ln Z_n &= \frac{1}{Z_n}(1-w)\mathcal{N}(\mathbf{x}_n|\mathbf{m}^{\backslash n},(v^{\backslash n}+1)\mathbf{I}) \\ &\qquad \left[\frac{1}{2(v^{\backslash n}+1)^2}\|\mathbf{x}_n-\mathbf{m}^{\backslash n}\|^2 - \frac{D}{2(v^{\backslash n}+1)}\right].\end{aligned}$$

Next we note that the variance is given by

$$v\mathbf{I} = \mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{\theta}]\mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}]$$

and so, taking the trace, we have

$$Dv = \mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}] - \mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}]\mathbb{E}[\boldsymbol{\theta}]$$

where D is the dimensionality of $\boldsymbol{\theta}$. Combining the above results we obtain (10.218).

Chapter 11 Sampling Methods

11.1 Since the samples are independent, for the mean, we have

$$\mathbb{E}\left[\widehat{f}\right] = \frac{1}{L}\sum_{l=1}^{L}\int f(z^{(l)})p(z^{(l)})\,\mathrm{d}z^{(l)} = \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[f\right] = \mathbb{E}\left[f\right].$$

Using this together with (1.38) and (1.39), for the variance, we have

$$\begin{aligned}\mathrm{var}\left[\widehat{f}\right] &= \mathbb{E}\left[\left(\widehat{f}-\mathbb{E}\left[\widehat{f}\right]\right)^2\right] \\ &= \mathbb{E}\left[\widehat{f}^2\right] - \mathbb{E}\left[f\right]^2.\end{aligned}$$

Now note

$$\begin{aligned}\mathbb{E}\left[f(z^{(k)}), f(z^{(m)})\right] &= \begin{cases} \mathrm{var}[f] + \mathbb{E}[f^2] & \text{if } n = k, \\ \mathbb{E}[f^2] & \text{otherwise,} \end{cases} \\ &= \mathbb{E}[f^2] + \delta_{mk}\mathrm{var}[f],\end{aligned}$$

where we again exploited the fact that the samples are independent.

Hence

$$\begin{aligned}\mathrm{var}\left[\widehat{f}\right] &= \mathbb{E}\left[\frac{1}{L}\sum_{m=1}^{L} f(z^{(m)})\frac{1}{L}\sum_{k=1}^{L} f(z^{(k)})\right] - \mathbb{E}[f]^2 \\ &= \frac{1}{L^2}\sum_{m=1}^{L}\sum_{k=1}^{L}\left\{\mathbb{E}[f^2] + \delta_{mk}\mathrm{var}[f]\right\} - \mathbb{E}[f]^2 \\ &= \frac{1}{L}\mathrm{var}[f] \\ &= \frac{1}{L}\mathbb{E}\left[(f - \mathbb{E}[f])^2\right].\end{aligned}$$

11.2 From (1.27) we have,

$$p_y(y) = p_z(h(y))\,|h'(y)|\,.$$

Differentiating (11.6) w.r.t. y and using the fact that $p_z(h(y)) = 1$, we see that

$$p_y(y) = p(y).$$

11.3 Using the standard integral

$$\int \frac{1}{a^2 + u^2}\,\mathrm{d}u = \frac{1}{a}\tan^{-1}\left(\frac{u}{a}\right) + C$$

where C is a constant, we can integrate the r.h.s. of (11.8) to obtain

$$z = h(y) = \int_{-\infty}^{y} p(\widehat{y})\,\mathrm{d}\widehat{y} = \frac{1}{\pi}\tan^{-1}(y) + \frac{1}{2}$$

where we have chosen the constant $C = 1/2$ to ensure that the range of the cumulative distribution function is $[0, 1]$.

Thus the required transformation function becomes

$$y = h^{-1}(z) = \tan\left(\pi\left(z - \frac{1}{2}\right)\right).$$

11.4 We need to calculate the determinant of the Jacobian

$$\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}.$$

In doing so, we will find it helpful to make use of intermediary variables in polar coordinates

$$\theta = \tan^{-1}\frac{z_2}{z_1} \tag{299}$$

$$r^2 = z_1^2 + z_2^2 \tag{300}$$

from which it follows that

$$z_1 = r\cos\theta \tag{301}$$

$$z_2 = r\sin\theta. \tag{302}$$

From (301) and (302) we have

$$\frac{\partial(z_1, z_2)}{\partial(r,\theta)} = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix}$$

and thus

$$\left|\frac{\partial(z_1, z_2)}{\partial(r,\theta)}\right| = r(\cos^2\theta + \sin^2\theta) = r. \tag{303}$$

From (11.10), (11.11) and (300)–(302) we have

$$y_1 = z_1\left(\frac{-2\ln r^2}{r^2}\right)^{1/2} = \left(-2\ln r^2\right)^{1/2}\cos\theta \tag{304}$$

$$y_2 = z_2\left(\frac{-2\ln r^2}{r^2}\right)^{1/2} = \left(-2\ln r^2\right)^{1/2}\sin\theta \tag{305}$$

which give

$$\frac{\partial(y_1, y_2)}{\partial(r,\theta)} = \begin{pmatrix} -2\cos\theta\left(-2\ln r^2\right)^{-1/2} r^{-1} & -2\sin\theta\left(-2\ln r^2\right)^{-1/2} r^{-1} \\ -\sin\theta\left(-2\ln r^2\right)^{1/2} & \cos\theta\left(-2\ln r^2\right)^{1/2} \end{pmatrix}$$

and thus

$$\left|\frac{\partial(r,\theta)}{\partial(y_1, y_2)}\right| = \left|\frac{\partial(y_1, y_2)}{\partial(r,\theta)}\right|^{-1} = \left(-2r^{-1}(\cos^2\theta + \sin^2\theta)\right)^{-1} = -\frac{r}{2}.$$

Combining this with (303), we get

$$\begin{aligned} \left|\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}\right| &= \left|\frac{\partial(z_1, z_2)}{\partial(r,\theta)}\frac{\partial(r,\theta)}{\partial(y_1, y_2)}\right| \\ &= \left|\frac{\partial(z_1, z_2)}{\partial(r,\theta)}\right|\left|\frac{\partial(r,\theta)}{\partial(y_1, y_2)}\right| = -\frac{r^2}{2} \end{aligned} \tag{306}$$

However, we only retain the absolute value of this, since both sides of (11.12) must be non-negative. Combining this with

$$p(z_1, z_2) = \frac{1}{\pi}$$

which follows from the fact that z_1 and z_2 are uniformly distributed on the unit circle, we can rewrite (11.12) as

$$p(y_1, y_2) = \frac{1}{2\pi} r^2. \tag{307}$$

By squaring the left- and rightmost sides of (304) and (305), adding up the results and rearranging, we see that

$$r^2 = \exp\left(-\frac{1}{2}\left(y_1^2 + y_2^2\right)\right)$$

which toghether with (307) give (11.12).

11.5 Since $\mathbb{E}[\mathbf{z}] = \mathbf{0}$,

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\boldsymbol{\mu} + \mathbf{L}\mathbf{z}] = \boldsymbol{\mu}.$$

Similarly, since $\mathbb{E}\left[\mathbf{z}\mathbf{z}^{\mathrm{T}}\right] = \mathbf{I}$,

$$\begin{aligned}
\operatorname{cov}[\mathbf{y}] &= \mathbb{E}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] - \mathbb{E}[\mathbf{y}]\,\mathbb{E}\left[\mathbf{y}^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[(\boldsymbol{\mu} + \mathbf{L}\mathbf{z})\,(\boldsymbol{\mu} + \mathbf{L}\mathbf{z})^{\mathrm{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \\
&= \mathbf{L}\mathbf{L}^{\mathrm{T}} \\
&= \boldsymbol{\Sigma}.
\end{aligned}$$

11.6 The probability of acceptance follows directly from the mechanism used to accept or reject the sample. The probability of a sample $\mathbf{z}$ being accepted equals the probability of a sample u, drawn uniformly from the interval $[0, kq(\mathbf{z})]$, being less than or equal to a value $\widetilde{p}(\mathbf{z}) \leqslant kq(\mathbf{z})$, and is given by is given by

$$p(\text{acceptance}|\mathbf{z}) = \int_0^{\widetilde{p}(\mathbf{z})} \frac{1}{kq(\mathbf{z})}\,\mathrm{d}u = \frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})}.$$

Therefore, the probability of drawing a sample, $\mathbf{z}$, is

$$q(\mathbf{z})p(\text{acceptance}|\mathbf{z}) = q(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})} = \frac{\widetilde{p}(\mathbf{z})}{k}. \tag{308}$$

Integrating both sides w.r.t. $\mathbf{z}$, we see that $kp(\text{acceptance}) = Z_p$, where

$$Z_p = \int \widetilde{p}(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Combining this with (308) and (11.13), we obtain

$$\frac{q(\mathbf{z})p(\text{acceptance}|\mathbf{z})}{p(\text{acceptance})} = \frac{1}{Z_p}\widetilde{p}(\mathbf{z}) = p(\mathbf{z})$$

as required.

11.7 **NOTE**: In PRML, the roles of y and z in the text of the exercise should be swapped in order to be consistent with the notation used in Section 11.1.2, including (11.16); this is opposite to the notation used in Section 11.1.1.

We will suppose that y has a uniform distribution on $[0, 1]$ and we seek the distribution of z, which is derived from

$$z = b \tan y + c.$$

From (11.5), we have

$$q(z) = p(y) \left| \frac{\mathrm{d}y}{\mathrm{d}z} \right| \tag{309}$$

From the inverse transformation

$$y = \tan^{-1}(u(z)) = \tan^{-1}\left(\frac{z-c}{b}\right)$$

where we have implicitly defined $u(z)$, we see that

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}z} &= \frac{\mathrm{d}}{\mathrm{d}u} \tan^{-1}(u) \frac{\mathrm{d}u}{\mathrm{d}z} \\ &= \frac{1}{1+u^2} \frac{\mathrm{d}u}{\mathrm{d}z} \\ &= \frac{1}{1+(z-c)^2/b^2} \frac{1}{b}. \end{aligned}$$

Substituting this into (309), using the fact that $p(y) = 1$ and finally absorbing the factor $1/b$ into k, we obtain (11.16). .

11.8 **NOTE**: In PRML, equation (11.17) and the following end of the sentence need to modified as follows:

$$q(z) = k_i \lambda_i \exp\left\{-\lambda_i \left(z - z_i\right)\right\} \qquad \widehat{z}_{i-1,i} < z \leqslant \widehat{z}_{i,i+1}$$

where $\widehat{z}_{i-1,i}$ is the point of intersection of the tangent lines at z_{i-1} and z_i, λ_i is the slope of the tangent at z_i and k_i accounts for the corresponding offset.

We start by determining $\widetilde{q}(z)$ with coefficients $\widetilde{k}_i$, such that $\widetilde{q}(z) \geqslant p(z)$ everywhere. From Figure 11.6, we see that

$$\widetilde{q}(z_i) = p(z_i)$$

and thus, from (11.17),

$$\begin{aligned} \widetilde{q}(z_i) &= \widetilde{k}_i \lambda_i \exp\left(-\lambda_i (z_i - z_i)\right) \\ &= \widetilde{k}_i \lambda_i = p(z_i). \end{aligned} \tag{310}$$

Next we compute the normalization constant for q in terms of k_i,

$$\begin{aligned} Z_q &= \int \widetilde{q}(z)\,\mathrm{d}z \\ &= \sum_{i=1}^{K} \widetilde{k}_i \lambda_i \int_{\widehat{z}_{i-1,i}}^{\widehat{z}_{i,i+1}} \exp\left(-\lambda_i(z-z_i)\right)\,\mathrm{d}z \\ &= \sum_{i=1}^{K} k_i \end{aligned} \tag{311}$$

where K denotes the number of grid points and

$$\begin{aligned} k_i &= \widetilde{k}_i \lambda_i \int_{\widehat{z}_{i-1,i}}^{\widehat{z}_{i,i+1}} \exp\left(-\lambda_i(z-z_i)\right)\,\mathrm{d}z \\ &= \widetilde{k}_i \left(\exp\left\{-\lambda_i\left(\widehat{z}_{i-1,i}-z_i\right)\right\} - \exp\left\{-\lambda_i\left(\widehat{z}_{i,i+1}-z_i\right)\right\}\right). \end{aligned} \tag{312}$$

Note that $\widehat{z}_{0,1}$ and $\widehat{z}_{K,K+1}$ equal the lower and upper limits on z, respectively, or $-\infty/+\infty$ where no such limits exist.

11.9 **NOTE**: See correction detailed in Solution 11.8

To generate a sample from $q(z)$, we first determine the segment of the envelope function from which the sample will be generated. The probability mass in segment i is given from (311) as k_i/Z_q. Hence we draw v from $U(v|0,1)$ and obtain

$$i = \begin{cases} 1 & \text{if } v \leqslant k_1/Z_q \\ m & \text{if } \sum_{j=1}^{m-1} k_j/Z_q < v \leqslant \sum_{j=1}^{m} k_j/Z_q, \quad 1 < m < K \\ K & \text{otherwise.} \end{cases}$$

Next, we can use the techniques of Section 11.1.1 to sample from the exponential distribution corresponding to the chosen segment i. We must, however, take into account that we now only want to sample from a finite interval of the exponential distribution and so the lower limit in (11.6) will be $\widehat{z}_{i-1,i}$. If we consider the uniformly distributed variable w, $U(w|0,1)$, (11.6), (310) and (311) give

$$\begin{aligned} w &= h(z) = \int_{\widehat{z}_{i-1,i}}^{z} q(\widetilde{z})\,\mathrm{d}\widetilde{z} \\ &= \frac{\widetilde{k}_i}{k_i}\lambda_i \exp\left(\lambda_i z_i\right) \int_{\widehat{z}_{i-1,i}}^{z} \exp\left(-\lambda_i \widetilde{z}\right)\,\mathrm{d}\widetilde{z} \\ &= \frac{\widetilde{k}_i}{k_i} \exp\left(\lambda_i z_i\right)\left[\exp\left(-\lambda_i \widehat{z}_{i-1,i}\right) - \exp\left(-\lambda_i z\right)\right]. \end{aligned}$$

Thus, by drawing a sample $w^\star$ and transforming it according to

$$\begin{aligned} z^\star &= \frac{1}{\lambda_i}\ln\left[w^\star\frac{k_i}{\widetilde{k}_i\exp\left(\lambda_i z_i\right)} - \exp\left(-\lambda_i\widehat{z}_{i-1,i}\right)\right] \\ &= \frac{1}{\lambda_i}\ln\left[w^\star\left(\exp\left\{-\lambda_i\widehat{z}_{i-1,i}\right\} - \exp\left\{-\lambda_i\widehat{z}_{i,i+1}\right\}\right) - \exp\left(-\lambda_i\widehat{z}_{i-1,i}\right)\right] \end{aligned}$$

where we have used (312), we obtain a sample from $q(z)$.

11.10 **NOTE**: In PRML, "$z^{(1)} = 0$" should be "$z^{(0)} = 0$" on the first line following Equation (11.36)

From (11.34)–(11.36) and the fact that $\mathbb{E}\left[z^{(\tau)}\right] = 0$ for all values of τ, we have

$$\begin{aligned} \mathbb{E}_{z^{(\tau)}}\left[\left(z^{(\tau)}\right)^2\right] &= 0.5\,\mathbb{E}_{z^{(\tau-1)}}\left[\left(z^{(\tau-1)}\right)^2\right] + 0.25\,\mathbb{E}_{z^{(\tau-1)}}\left[\left(z^{(\tau-1)}+1\right)^2\right] \\ &\quad +0.25\,\mathbb{E}_{z^{(\tau-1)}}\left[\left(z^{(\tau-1)}-1\right)^2\right] \\ &= \mathbb{E}_{z^{(\tau-1)}}\left[\left(z^{(\tau-1)}\right)^2\right] + \frac{1}{2}. \end{aligned} \tag{313}$$

With $z^{(0)} = 0$ specified in the text following (11.36), (313) gives

$$\mathbb{E}_{z^{(1)}}\left[\left(z^{(1)}\right)^2\right] = \mathbb{E}_{z^{(0)}}\left[\left(z^{(0)}\right)^2\right] + \frac{1}{2} = \frac{1}{2}.$$

Assuming that

$$\mathbb{E}_{z^{(k)}}\left[\left(z^{(k)}\right)^2\right] = \frac{k}{2}$$

(313) immediatly gives

$$\mathbb{E}_{z^{(k+1)}}\left[\left(z^{(k+1)}\right)^2\right] = \frac{k}{2} + \frac{1}{2} = \frac{k+1}{2}$$

and thus

$$\mathbb{E}_{z^{(\tau)}}\left[\left(z^{(\tau)}\right)^2\right] = \frac{\tau}{2}.$$

11.11 This follows from the fact that in Gibbs sampling, we sample a single variable, z_k, at the time, while all other variables, $\{z_i\}_{i\neq k}$, remain unchanged. Thus, $\{z_i'\}_{i\neq k} = \{z_i\}_{i\neq k}$ and we get

$$\begin{aligned} p^\star(\mathbf{z})T(\mathbf{z},\mathbf{z}') &= p^\star(z_k,\{z_i\}_{i\neq k})p^\star(z_k'|\{z_i\}_{i\neq k}) \\ &= p^\star(z_k|\{z_i\}_{i\neq k})p^\star(\{z_i\}_{i\neq k})p^\star(z_k'|\{z_i\}_{i\neq k}) \\ &= p^\star(z_k|\{z_i'\}_{i\neq k})p^\star(\{z_i'\}_{i\neq k})p^\star(z_k'|\{z_i'\}_{i\neq k}) \\ &= p^\star(z_k|\{z_i'\}_{i\neq k})p^\star(z_k',\{z_i'\}_{i\neq k}) \\ &= p^\star(\mathbf{z}')T(\mathbf{z}',\mathbf{z}), \end{aligned}$$

where we have used the product rule together with $T(\mathbf{z},\mathbf{z}') = p^\star(z_k'|\{z_i\}_{i\neq k})$.

11.12 Gibbs sampling is *not* ergodic w.r.t. the distribution shown in Figure 11.15, since the two regions of non-zero probability do not overlap when projected onto either the z_1-ot the z_2-axis. Thus, as the initial sample will fall into one and only one of the two regions, all subsequent samples will also come from that region. However, had the initial sample fallen into the other region, the Gibbs sampler would have remained in that region. Thus, the corresponding Markov chain would have two stationary distributions, which is counter to the definition of the equilibrium distribution.

11.13 The joint distribution over x, μ and τ can be written

$$p(x, \mu, \tau|\mu_0, s_0, a, b) = \mathcal{N}\left(x|\mu, \tau^{-1}\right) \mathcal{N}\left(\mu|\mu_0, s_0\right) \text{Gam}\left(\tau|a, b\right).$$

From Bayes' theorem we see that

$$p(\mu|x, \tau, \mu_0, s_0) \propto \mathcal{N}\left(x|\mu, \tau^{-1}\right) \mathcal{N}\left(\mu|\mu_0, s_0\right)$$

which, from (2.113)–(2.117), is also a Gaussian,

$$\mathcal{N}\left(\mu|\widehat{\mu}, \widehat{s}\right)$$

with parameters

$$\begin{aligned} \widehat{s}^{-1} &= s_0^{-1} + \tau \\ \widehat{\mu} &= \widehat{s}\left(\tau x + s_0^{-1}\mu_0\right). \end{aligned}$$

Similarly,

$$p(\tau|x, \mu, a, b) \propto \mathcal{N}\left(x|\mu, \tau^{-1}\right) \text{Gam}\left(\tau|a, b\right)$$

which, we know from Section 2.3.6, is a gamma distribution

$$\text{Gam}\left(\tau|\widehat{a}, \widehat{b}\right)$$

with parameters

$$\begin{aligned} \widehat{a} &= a + \frac{1}{2} \\ \widehat{b} &= b + \frac{1}{2}\left(x - \mu\right)^2. \end{aligned}$$

11.14 **NOTE**: In PRML, α_i^2 should be α^2 in the last term on the r.h.s. of (11.50).

If we take the expectation of (11.50), we obtain

$$\begin{aligned} \mathbb{E}\left[z_i'\right] &= \mathbb{E}\left[\mu_i + \alpha\left(z_i - \mu_i\right) + \sigma_i\left(1 - \alpha^2\right)^{1/2}\nu\right] \\ &= \mu_i + \alpha\left(\mathbb{E}\left[z_i\right] - \mu_i\right) + \sigma_i\left(1 - \alpha^2\right)^{1/2}\mathbb{E}\left[\nu\right] \\ &= \mu_i. \end{aligned}$$

Now we can use this together with (1.40) to compute the variance of z_i',

$$\begin{aligned}
\text{var}\left[z_i'\right] &= \mathbb{E}\left[\left(z_i'\right)^2\right] - \mathbb{E}\left[z_i'\right]^2 \\
&= \mathbb{E}\left[\left(\mu_i + \alpha\left(z_i - \mu_i\right) + \sigma_i\left(1-\alpha^2\right)^{1/2}\nu\right)^2\right] - \mu_i^2 \\
&= \alpha^2\mathbb{E}\left[\left(z_i - \mu_i\right)^2\right] + \sigma_i^2\left(1-\alpha^2\right)\mathbb{E}\left[\nu^2\right] \\
&= \sigma_i^2
\end{aligned}$$

where we have used the first and second order moments of z_i and ν.

11.15 Using (11.56), we can differentiate (11.57), yielding

$$\frac{\partial H}{\partial r_i} = \frac{\partial K}{\partial r_i} = r_i$$

and thus (11.53) and (11.58) are equivalent.

Similarly, differentiating (11.57) w.r.t. z_i we get

$$\frac{\partial H}{\partial z_i} = \frac{\partial E}{\partial z_i},$$

and from this, it is immediately clear that (11.55) and (11.59) are equivalent.

11.16 From the product rule we know that

$$p(\mathbf{r}|\mathbf{z}) \propto p(\mathbf{r}, \mathbf{z}).$$

Using (11.56) and (11.57) to rewrite (11.63) as

$$\begin{aligned}
p(\mathbf{z}, \mathbf{r}) &= \frac{1}{Z_H}\exp\left(-H(\mathbf{z}, \mathbf{r})\right) \\
&= \frac{1}{Z_H}\exp\left(-E(\mathbf{z}) - K(\mathbf{r})\right) \\
&= \frac{1}{Z_H}\exp\left(-\frac{1}{2}\|\mathbf{r}\|^2\right)\exp\left(-E(\mathbf{z})\right).
\end{aligned}$$

Thus we see that $p(\mathbf{z}, \mathbf{r})$ is Gaussian w.r.t. $\mathbf{r}$ and hence $p(\mathbf{r}|\mathbf{z})$ will be Gaussian too.

11.17 **NOTE**: In the 1^{st} printing of PRML, there are sign errors in equations (11.68) and (11.69). In both cases, the sign of the argument to the exponential forming the second argument to the $\min$-function should be changed.

First we note that, if $H(\mathcal{R}) = H(\mathcal{R}')$, then the detailed balance clearly holds, since in this case, (11.68) and (11.69) are identical.

Otherwise, we either have $H(\mathcal{R}) > H(\mathcal{R}')$ or $H(\mathcal{R}) < H(\mathcal{R}')$. We consider the former case, for which (11.68) becomes

$$\frac{1}{Z_H}\exp(-H(\mathcal{R}))\delta V\frac{1}{2},$$

since the min-function will return 1. (11.69) in this case becomes

$$\frac{1}{Z_H}\exp(-H(\mathcal{R}'))\delta V\frac{1}{2}\exp(H(\mathcal{R}')-H(\mathcal{R}))=\frac{1}{Z_H}\exp(-H(\mathcal{R}))\delta V\frac{1}{2}.$$

In the same way it can be shown that both (11.68) and (11.69) equal

$$\frac{1}{Z_H}\exp(-H(\mathcal{R}'))\delta V\frac{1}{2}$$

when $H(\mathcal{R})<H(\mathcal{R}')$.

Chapter 12 Continuous Latent Variables

12.1 Suppose that the result holds for projection spaces of dimensionality M. The $M+1$ dimensional principal subspace will be defined by the M principal eigenvectors $\mathbf{u}_1,\ldots,\mathbf{u}_M$ together with an additional direction vector $\mathbf{u}_{M+1}$ whose value we wish to determine. We must constrain $\mathbf{u}_{M+1}$ such that it cannot be linearly related to $\mathbf{u}_1,\ldots,\mathbf{u}_M$ (otherwise it will lie in the M-dimensional projection space instead of defining an $M+1$ independent direction). This can easily be achieved by requiring that $\mathbf{u}_{M+1}$ be orthogonal to $\mathbf{u}_1,\ldots,\mathbf{u}_M$, and these constraints can be enforced using Lagrange multipliers $\eta_1,\ldots,\eta_M$.

Following the argument given in section 12.1.1 for $\mathbf{u}_1$ we see that the variance in the direction $\mathbf{u}_{M+1}$ is given by $\mathbf{u}_{M+1}^{\mathrm{T}}\mathbf{S}\mathbf{u}_{M+1}$. We now maximize this using a Lagrange multiplier λ_{M+1} to enforce the normalization constraint $\mathbf{u}_{M+1}^{\mathrm{T}}\mathbf{u}_{M+1}=1$. Thus we seek a maximum of the function

$$\mathbf{u}_{M+1}^{\mathrm{T}}\mathbf{S}\mathbf{u}_{M+1}+\lambda_{M+1}\left(1-\mathbf{u}_{M+1}^{\mathrm{T}}\mathbf{u}_{M+1}\right)+\sum_{i=1}^{M}\eta_i\mathbf{u}_{M+1}^{\mathrm{T}}\mathbf{u}_i.$$

with respect to $\mathbf{u}_{M+1}$. The stationary points occur when

$$0=2\mathbf{S}\mathbf{u}_{M+1}-2\lambda_{M+1}\mathbf{u}_{M+1}+\sum_{i=1}^{M}\eta_i\mathbf{u}_i.$$

Left multiplying with $\mathbf{u}_j^{\mathrm{T}}$, and using the orthogonality constraints, we see that $\eta_j=0$ for $j=1,\ldots,M$. We therefore obtain

$$\mathbf{S}\mathbf{u}_{M+1}=\lambda_{M+1}\mathbf{u}_{M+1}$$

and so $\mathbf{u}_{M+1}$ must be an eigenvector of $\mathbf{S}$ with eigenvalue $\mathbf{u}_{M+1}$. The variance in the direction $\mathbf{u}_{M+1}$ is given by $\mathbf{u}_{M+1}^{\mathrm{T}}\mathbf{S}\mathbf{u}_{M+1}=\lambda_{M+1}$ and so is maximized by choosing $\mathbf{u}_{M+1}$ to be the eigenvector having the largest eigenvalue amongst those not previously selected. Thus the result holds also for projection spaces of dimensionality $M+1$, which completes the inductive step. Since we have already shown this result explicitly for $M=1$ if follows that the result must hold for any $M\leqslant D$.

12.2 Using the result (C.24) we can set the derivative of $\widetilde{J}$ with respect to $\widehat{\mathbf{U}}$ to zero to obtain

$$0 = (\mathbf{S}^{\mathrm{T}} + \mathbf{S})\widehat{\mathbf{U}} - \widehat{\mathbf{U}}(\mathbf{H}^{\mathrm{T}} + \mathbf{H}).$$

We note that $\mathbf{S}$ is symmetric so that $\mathbf{S}^{\mathrm{T}} = \mathbf{S}$. Similarly we can choose $\mathbf{H}$ to be symmetric without loss of generality since any non-symmetric component would cancel from the expression for $\widetilde{J}$ since the latter involves a trace of $\mathbf{H}$ times a symmetric matrix. (See Exercise 1.14 and its solution.) Thus we have

$$\mathbf{S}\widehat{\mathbf{U}} = \widehat{\mathbf{U}}\mathbf{H}.$$

Clearly one solution is take the columns of $\widehat{\mathbf{U}}$ to be eigenvectors of $\mathbf{S}$. To discuss the general solution, consider the eigenvector equation for $\mathbf{H}$ given by

$$\mathbf{H}\boldsymbol{\Psi} = \boldsymbol{\Psi}\mathbf{L}.$$

Since $\mathbf{H}$ is a symmetric matrix its eigenvectors can be chosen to be a complete orthonormal set in the $(D - M)$-dimensional space, and $\mathbf{L}$ will be a diagonal matrix containing the corresponding eigenvalues, with $\boldsymbol{\Psi}$ a $(D-M)\times(D-M)$-dimensional orthogonal matrix satisfying $\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} = \mathbf{I}$.

If we right multiply the eigenvector equation for $\mathbf{S}$ by $\boldsymbol{\Psi}$ we obtain

$$\mathbf{S}\widehat{\mathbf{U}}\boldsymbol{\Psi} = \widehat{\mathbf{U}}\mathbf{H}\boldsymbol{\Psi} = \widehat{\mathbf{U}}\boldsymbol{\Psi}\mathbf{L}$$

and defining $\widetilde{\mathbf{U}} = \widehat{\mathbf{U}}\boldsymbol{\Psi}$ we obtain

$$\mathbf{S}\widetilde{\mathbf{U}} = \widetilde{\mathbf{U}}\mathbf{L}$$

so that the columns of $\widetilde{\mathbf{U}}$ are the eigenvectors of $\mathbf{S}$, and the elements of the diagonal matrix $\mathbf{L}$ are the corresponding eigenvalues.

Using the cyclic property of the trace, together with the orthogonality property $\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}$, the distortion function can be written

$$J = \mathrm{Tr}(\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{S}\widehat{\mathbf{U}}) = \mathrm{Tr}(\boldsymbol{\Psi}^{\mathrm{T}}\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{S}\widehat{\mathbf{U}}\boldsymbol{\Psi}) = \mathrm{Tr}(\widetilde{\mathbf{U}}\mathbf{S}\widetilde{\mathbf{U}}) = \mathrm{Tr}(\mathbf{L}).$$

Thus the distortion measure can be expressed in terms of the sum of the eigenvalues of $\mathbf{S}$ corresponding to the $(D-M)$ eigenvectors orthogonal to the principal subspace.

12.3 By left-multiplying both sides of (12.28) by $\mathbf{v}_i^{\mathrm{T}}$, we obtain

$$\frac{1}{N}\mathbf{v}_i^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{v}_i = \lambda_i\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i = \lambda_i$$

where we have used the fact that $\mathbf{v}_i$ is orthonormal. From this we see that

$$\left\|\mathbf{X}^{\mathrm{T}}\mathbf{v}_i\right\|^2 = N\lambda_i$$

from which we in turn see that $\mathbf{u}_i$ defined by (12.30) will have unit length.

12.4 Using the results of Section 8.1.4, the marginal distribution for this modified probabilistic PCA model can be written

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{m} + \boldsymbol{\mu}, \sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{W}).$$

If we now define new parameters

$$\begin{aligned} \widetilde{\mathbf{W}} &= \boldsymbol{\Sigma}^{1/2}\mathbf{W} \\ \widetilde{\boldsymbol{\mu}} &= \mathbf{W}\mathbf{m} + \boldsymbol{\mu} \end{aligned}$$

then we obtain a marginal distribution having the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\widetilde{\boldsymbol{\mu}}, \sigma^2\mathbf{I} + \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{W}}).$$

Thus any Gaussian form for the latent distribution therefore gives rise to a predictive distribution having the same functional form, and so for convenience we choose the simplest form, namely one with zero mean and unit covariance.

12.5 Since $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$,

$$p(\mathbf{y}|\mathbf{x}) = \delta(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})$$

i.e. a delta function at $\mathbf{A}\mathbf{x} + \mathbf{b}$. From the sum and product rules, we have

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y}, \mathbf{x})\,\mathrm{d}\mathbf{x} = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int \delta(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \end{aligned}$$

When $M = D$ and $\mathbf{A}$ is assumed to have full rank, we have

$$\mathbf{x} = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$$

and thus

$$\begin{aligned} p(\mathbf{y}) &= \mathcal{N}\left(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\right). \end{aligned}$$

When $M > D$, $\mathbf{y}$ will be strictly confined to a D-dimensional subspace and hence $p(\mathbf{y})$ will be singular. In this case we have

$$\mathbf{x} = \mathbf{A}^{-\mathrm{L}}(\mathbf{y} - \mathbf{b})$$

where $\mathbf{A}^{-\mathrm{L}}$ is the left inverse of $\mathbf{A}$ and thus

$$\begin{aligned} p(\mathbf{y}) &= \mathcal{N}\left(\mathbf{A}^{-\mathrm{L}}(\mathbf{y} - \mathbf{b})|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \left(\left(\mathbf{A}^{-\mathrm{L}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-\mathrm{L}}\right)^{-1}\right). \end{aligned}$$

The covariance matrix on the last line cannot be computed, but we can still compute $p(\mathbf{y})$, by using the corresponding precision matrix and constraining the density to be zero outside the column space of $\mathbf{A}$:

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \left(\left(\mathbf{A}^{-\mathrm{L}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-\mathrm{L}}\right)^{-1}\right)\delta\left(\mathbf{y}-\mathbf{A}\mathbf{A}^{-\mathrm{L}}\left(\mathbf{y}-\mathbf{b}\right)-\mathbf{b}\right).$$

Finally, when $M < D$, we can make use of (2.113)–(2.115) and set $\mathbf{L}^{-1} = \mathbf{0}$ in (2.114). While this means that $p(\mathbf{y}|\mathbf{x})$ is singular, the marginal distribution $p(\mathbf{y})$, given by (2.115), is non-singular as long as $\mathbf{A}$ and $\boldsymbol{\Sigma}$ are assumed to be of full rank.

12.6 Omitting the parameters, $\mathbf{W}$, $\boldsymbol{\mu}$ and σ, leaving only the stochastic variables $\mathbf{z}$ and $\mathbf{x}$, the graphical model for probabilistic PCA is identical with the the 'naive Bayes' model shown in Figure 8.24 in Section 8.2.2. Hence these two models exhibit the same independence structure.

12.7 From (2.59), the multivariate form of (2.270), (12.31) and (12.32), we get

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \mathbb{E}_{\mathbf{z}}\left[\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}|\mathbf{z}\right]\right]\\ &= \mathbb{E}_{\mathbf{z}}\left[\mathbf{W}\mathbf{z}+\boldsymbol{\mu}\right]\\ &= \boldsymbol{\mu}.\end{aligned}$$

Combining this with (2.63), the covariance formula corresponding to (2.271), (12.31) and (12.32), we get

$$\begin{aligned}\mathrm{cov}[\mathbf{x}] &= \mathbb{E}_{\mathbf{z}}\left[\mathrm{cov}_{\mathbf{x}}[\mathbf{x}|\mathbf{z}]\right] + \mathrm{cov}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{x}}[\mathbf{x}|\mathbf{z}]\right]\\ &= \mathbb{E}_{\mathbf{z}}\left[\sigma^2\mathbf{I}\right] + \mathrm{cov}_{\mathbf{z}}\left[\mathbf{W}\mathbf{z}+\boldsymbol{\mu}\right]\\ &= \sigma^2\mathbf{I} + \mathbb{E}_{\mathbf{z}}\left[\left(\mathbf{W}\mathbf{z}+\boldsymbol{\mu}-\mathbb{E}_{\mathbf{z}}\left[\mathbf{W}\mathbf{z}+\boldsymbol{\mu}\right]\right)\left(\mathbf{W}\mathbf{z}+\boldsymbol{\mu}-\mathbb{E}_{\mathbf{z}}\left[\mathbf{W}\mathbf{z}+\boldsymbol{\mu}\right]\right)^{\mathrm{T}}\right]\\ &= \sigma^2\mathbf{I} + \mathbb{E}_{\mathbf{z}}\left[\mathbf{W}\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\right]\\ &= \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}}.\end{aligned}$$

12.8 **NOTE**: In the 1st printing of PRML, equation (12.42) contains a mistake; the covariance on the r.h.s. should be $\sigma^2\mathbf{M}^{-1}$.

By matching (12.31) with (2.113) and (12.32) with (2.114), we have from (2.116) and (2.117) that

$$\begin{aligned}p(\mathbf{z}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{z}|(\mathbf{I}+\sigma^{-2}\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\sigma^{-2}\mathbf{I}(\mathbf{x}-\boldsymbol{\mu}), (\mathbf{I}+\sigma^{-2}\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\right)\\ &= \mathcal{N}\left(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}\right),\end{aligned}$$

where we have also used (12.41).

12.9 By expanding the square in the last term of (12.43) and then making use of results from Appendix C, we can calculate the derivative w.r.t. $\boldsymbol{\mu}$ and set this equal to zero, yielding

$$-N\mathbf{C}^{-1}\boldsymbol{\mu} + \mathbf{C}^{-1}\sum_{n=1}^{N}\mathbf{x}_n = \mathbf{0}. \tag{314}$$

Rearrangning this and making use of (12.1), we get

$$\boldsymbol{\mu} = \overline{\mathbf{x}}.$$

12.10 Using results from Appendix C, we can differentiate the r.h.s. of (314) w.r.t. $\boldsymbol{\mu}$, giving

$$-N\mathbf{C}^{-1}.$$

If $\sigma^2 > 0$, $\mathbf{C}$ will be positive definite, in which case $\mathbf{C}^{-1}$ is also positive definite, and hence the log likelihood function will be concave with a unique maximum at $\boldsymbol{\mu}$.

12.11 Taking $\sigma^2 \to 0$ in (12.41) and substituting into (12.48) we obtain the posterior mean for probabilistic PCA in the form

$$(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}})^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{x} - \overline{\mathbf{x}}).$$

Now substitute for $\mathbf{W}_{\mathrm{ML}}$ using (12.45) in which we take $\mathbf{R} = \mathbf{I}$ for compatibility with conventional PCA. Using the orthogonality property $\mathbf{U}_M^{\mathrm{T}}\mathbf{U}_M = \mathbf{I}$ and setting $\sigma^2 = 0$, this reduces to

$$\mathbf{L}^{-1/2}\mathbf{U}_M^{\mathrm{T}}(\mathbf{x} - \overline{\mathbf{x}})$$

which is the orthogonal projection is given by the conventional PCA result (12.24).

12.12 For $\sigma^2 > 0$ we can show that the projection is shifted towards the origin of latent space by showing that the magnitude of the latent space vector is reduced compared to the $\sigma^2 = 0$ case. The orthogonal projection is given by

$$\mathbf{z}_{\mathrm{orth}} = \mathbf{L}_M^{-1/2}\mathbf{U}_M^{\mathrm{T}}(\mathbf{x} - \overline{\mathbf{x}})$$

where L_M and U_M are defined as in (12.45). The posterior mean projection is given by (12.48) and so the difference between the squared magnitudes of each of these is given by

$$\begin{aligned}
&\|\mathbf{z}_{\mathrm{orth}}\|^2 - \|\mathbb{E}[\mathbf{z}|\mathbf{x}]\|^2 \\
&\quad = (\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}}\left(\mathbf{U}_M\mathbf{L}_M^{-1/2}\mathbf{U}_M - \mathbf{W}_{\mathrm{ML}}\mathbf{M}^{-1}\mathbf{M}^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\right)(\mathbf{x} - \overline{\mathbf{x}}) \\
&\quad = (\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}}\mathbf{U}_M\left\{\mathbf{L}^{-1} - (\mathbf{L} + \sigma^2\mathbf{I})^{-1}\right\}\mathbf{U}_M^{\mathrm{T}}(\mathbf{x} - \overline{\mathbf{x}})
\end{aligned}$$

where we have use (12.41), (12.45) and the fact that $\mathbf{L}$ and $\mathbf{M}$ are symmetric. The term in curly braces on the last line is diagonal and has elements $\sigma^2/\lambda_i(\lambda_i + \sigma^2)$ which are all positive. Thus the matrix is positive definite and so the contraction with the vector $\mathbf{U}_M(\mathbf{x} - \overline{\mathbf{x}})$ must be positive, and so there is a positive (non-zero) shift towards the origin.

12.13 Substituting the r.h.s. of (12.48) for $\mathbb{E}[\mathbf{z}|\mathbf{x}]$ in (12.94), we obtain,

$$\widetilde{\mathbf{x}}_n = \mathbf{W}_{\mathrm{ML}}\left(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}\right)^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{x}_n - \overline{\mathbf{x}}).$$

From (12.45) we see that

$$\left(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}\right)^{-1} = \left(\mathbf{L}_M - \sigma^2\mathbf{I}\right)^{-1}$$

and so

$$\widetilde{\mathbf{x}}_n = \mathbf{U}_M\mathbf{U}_M^{\mathrm{T}}(\mathbf{x}_n - \overline{\mathbf{x}}).$$

This is the reconstruction of $\mathbf{x}_n - \overline{\mathbf{x}}$ using the M eigenvectors corresponding to the M largest eigenvalues, which we know from Section 12.1.2 minimizes the least squares projection cost (12.11).

12.14 If we substitute $D-1$ for M in (12.51), we get

$$\begin{aligned} D(D-1) + 1 - \frac{(D-1)((D-1)-1)}{2} &= \frac{2D^2 - 2D + 2 - D^2 + 3D - 2}{2} \\ &= \frac{D^2 + D}{2} = \frac{D(D+1)}{2} \end{aligned}$$

as required. Setting $M = 0$ in (12.51) given the value 1 for the number of parameters in $\mathbf{C}$, corresponding to the scalar variance parameter, σ^2.

12.15 **NOTE**: In PRML, a term $M/2\ln(2\pi)$ is missing from the summand on the r.h.s. of (12.53). However, this is only stated here for completeness as it actually does not affect this solution.

Using standard derivatives together with the rules for matrix differentiation from Appendix C, we can compute the derivatives of (12.53) w.r.t. $\mathbf{W}$ and σ^2:

$$\frac{\partial}{\partial \mathbf{W}}\mathbb{E}[\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right)] = \sum_{n=1}^{N}\left\{\frac{1}{\sigma^2}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}} - \frac{1}{\sigma^2}\mathbf{W}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\right\}$$

and

$$\begin{aligned} \frac{\partial}{\partial \sigma^2}\mathbb{E}[\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right)] = \sum_{n=1}^{N}\Bigg\{ &\frac{1}{2\sigma^4}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}}\mathbf{W} \\ &+ \frac{1}{2\sigma^4}\|\mathbf{x}_n - \overline{\mathbf{x}}\|^2 - \frac{1}{\sigma^4}\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_n - \overline{\mathbf{x}}) - \frac{D}{2\sigma^2}\Bigg\} \end{aligned}$$

Setting these equal to zero and re-arranging we obtain (12.56) and (12.57), respectively.

12.16 We start by noting that the marginal likelihood factorizes over data points as well as the individual elements of the data points,

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) &= \int p(\mathbf{Z}) p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \mathbf{W}, \sigma^2)\, \mathrm{d}\mathbf{Z} \\ &= \prod_{n=1}^{N} \int p(\mathbf{z}_n) p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\mu}, \mathbf{W}, \sigma^2)\, \mathrm{d}\mathbf{z}_n \\ &= \prod_{n=1}^{N} \int p(\mathbf{z}_n) \prod_{i=1}^{D} \mathcal{N}(x_{ni}|\mathbf{w}_i \mathbf{z}_n + \mu_i, \sigma^2)\, \mathrm{d}\mathbf{z}_n \end{aligned} \tag{315}$$

where x_{ni} denotes the i^{th} element of $\mathbf{x}_n$, μ_i denotes the i^{th} element of $\boldsymbol{\mu}$ and $\mathbf{w}_i$ the i^{th} row of $\mathbf{W}$. If we assume that any missing values are missing at random (see page 441 of PRML), we can deal with these by integrating them out of (315). Let $\mathbf{x}_n^{\mathrm{o}}$ and $\mathbf{x}_n^{\mathrm{m}}$ denote the observed and missing parts of $\mathbf{x}_n$, respectively. Using this notation, we can rewrite (315) as

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) &= \prod_{n=1}^{N} \int p(\mathbf{z}_n) \prod_{x_{ni} \in \mathbf{x}_n^{\mathrm{o}}} \mathcal{N}(x_{ni}|\mathbf{w}_i \mathbf{z}_n + \mu_i, \sigma^2) \\ &\qquad \int \prod_{x_{nj} \in \mathbf{x}_n^{\mathrm{m}}} \mathcal{N}(x_{nj}|\mathbf{w}_j \mathbf{z}_n + \mu_j, \sigma^2)\, \mathrm{d}\mathbf{x}_n^{\mathrm{m}}\, \mathrm{d}\mathbf{z}_n \\ &= \prod_{n=1}^{N} \int p(\mathbf{z}_n) \prod_{x_{ni} \in \mathbf{x}_n^{\mathrm{o}}} \mathcal{N}(x_{ni}|\mathbf{w}_i \mathbf{z}_n + \mu_i, \sigma^2)\, \mathrm{d}\mathbf{z}_n \\ &= \prod_{n=1}^{N} p(\mathbf{x}_n^{\mathrm{o}}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2). \end{aligned}$$

Thus we are left with a 'reduced' marginal likelihood, where for each data point, $\mathbf{x}_n$, we only need to consider the observed elements, $\mathbf{x}_n^{\mathrm{o}}$.

Now we can derive an EM algorithm for finding the parameter values that maximizes this 'reduced' marginal likelihood. In doing so, we shall find it convenient to introduce indicator variables, ι_{ni}, such that $\iota_{ni} = 1$ if x_{ni} is observed and $\iota_{ni} = 0$ otherwise. This allows us to rewrite (12.32) as

$$p(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{D} \mathcal{N}(x_i|\mathbf{w}_i \mathbf{z} + \mu_i, \sigma^2)^{\iota_{ni}}$$

and the complete-data log likelihood as

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma) = \sum_{n=1}^{N} \left\{ \ln p(\mathbf{z}_n) + \sum_{i=1}^{D} \iota_{ni} \ln \mathcal{N}(x_{ni}|\mathbf{w}_i \mathbf{z}_n + \mu_i, \sigma^2) \right\}.$$

Following the path taken in Section 12.2.2, making use of (12.31) and taking the expectation w.r.t. the latent variables, we obtain

$$\mathbb{E}\left[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma)\right] = -\sum_{n=1}^{N}\left\{\frac{M}{2}\ln(2\pi) + \frac{1}{2}\mathrm{Tr}\left(\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right]\right)\right.$$
$$+\sum_{i=1}^{D}\iota_{ni}\left\{\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}(x_{ni}-\mu_{ni})^2 - \frac{1}{\sigma^2}\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\mathbf{w}_i^{\mathrm{T}}(x_{ni}-\mu_{ni})\right.$$
$$\left.\left.+\frac{1}{2\sigma^2}\mathrm{Tr}\left(\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right]\mathbf{w}_i^T\mathbf{w}_i\right)\right\}\right\}.$$

Taking the derivative w.r.t. μ_i and setting the result equal to zero, we obtain

$$\mu_i^{\mathrm{new}} = \frac{1}{\sum_{m=1}^{N}\iota_{mi}}\sum_{n=1}^{N}\iota_{ni}x_{ni}.$$

In the E step, we compute the sufficient statistics, which due to the altered form of $p(\mathbf{x}|\mathbf{z})$ now take slightly different shapes. Equation (12.54) becomes

$$\mathbb{E}[\mathbf{z}_n] = \mathbf{M}_n^{-1}\mathbf{W}_n^{\mathrm{T}}\mathbf{y}_n$$

where $\mathbf{y}_n$ is a vector containing the observed elements of $\mathbf{x}_n$ minus the corresponding elements of μ^{new}, $\mathbf{W}_n$ is a matrix formed by the rows of $\mathbf{W}$ corresponding to the observed elements of $\mathbf{x}_n$ and, accordingly, from (12.41)

$$\mathbf{M}_n = \mathbf{W}_n^{\mathrm{T}}\mathbf{W}_n + \sigma^2\mathbf{I}.$$

Similarly,

$$\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right] = \sigma^2\mathbf{M}_n^{-1} + \mathbb{E}[\mathbf{z}_n]\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}.$$

The M step is similar to the fully observed case, with

$$\begin{aligned}
\mathbf{W}_{\mathrm{new}} &= \left[\sum_{n=1}^{N}\mathbf{y}_n\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\right]\left[\sum_{n=1}^{N}\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right]\right]^{-1} \\
\sigma^2_{\mathrm{new}} &= \frac{1}{\sum_{n=1}^{N}\sum_{i=1}^{D}\iota_{ni}}\sum_{n=1}^{N}\sum_{i=1}^{D}\iota_{ni}\left\{(x_{ni}-\mu_i^{\mathrm{new}})^2\right. \\
&\quad -2\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\left(\mathbf{w}_i^{\mathrm{new}}\right)^{\mathrm{T}}\left(x_{ni}-\mu_i^{\mathrm{new}}\right) \\
&\quad \left.+\mathrm{Tr}\left(\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right]\left(\mathbf{w}_i^{\mathrm{new}}\right)^{\mathrm{T}}\mathbf{w}_i^{\mathrm{new}}\right)\right\}
\end{aligned}$$

where $\mathbf{w}_i^{\mathrm{new}}$ equals the i^{th} row $\mathbf{W}_{\mathrm{new}}$.

In the fully observed case, all $\iota_{ni} = 1$, $\mathbf{y}_n = \mathbf{x}_n$, $\mathbf{W}_n = \mathbf{W}$ and $\boldsymbol{\mu}^{\mathrm{new}} = \overline{\mathbf{x}}$, and hence we recover (12.54)–(12.57).

12.17 **NOTE**: In PRML, there are errors in equation (12.58) and the preceding text. In (12.58), $\widetilde{\mathbf{X}}$ should be $\widetilde{\mathbf{X}}^{\mathrm{T}}$ and in the preceding text we define $\mathbf{\Omega}$ to be a matrix of size $M \times N$ whose n^{th} column is given by the vector $\mathbb{E}[\mathbf{z}_n]$.

Setting the derivative of J with respect to $\boldsymbol{\mu}$ to zero gives

$$0 = -\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_n)$$

from which we obtain

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n - \frac{1}{N}\sum_{n=1}^{N}\mathbf{W}\mathbf{z}_n = \overline{\mathbf{x}} - \mathbf{W}\overline{\mathbf{z}}.$$

Back-substituting into J we obtain

$$J = \sum_{n=1}^{N}\|(\mathbf{x}_n - \overline{\mathbf{x}} - \mathbf{W}(\mathbf{z}_n - \overline{\mathbf{z}})\|^2.$$

We now define $\mathbf{X}$ to be a matrix of size $N \times D$ whose n^{th} row is given by the vector $\mathbf{x}_n - \overline{\mathbf{x}}$ and similarly we define $\mathbf{Z}$ to be a matrix of size $D \times M$ whose n^{th} row is given by the vector $\mathbf{z}_n - \overline{\mathbf{z}}$. We can then write J in the form

$$J = \mathrm{Tr}\left\{(\mathbf{X} - \mathbf{Z}\mathbf{W}^{\mathrm{T}})(\mathbf{X} - \mathbf{Z}\mathbf{W}^{\mathrm{T}})^{\mathrm{T}}\right\}.$$

Differentiating with respect to $\mathbf{Z}$ keeping $\mathbf{W}$ fixed gives rise to the PCA E-step (12.58). Similarly setting the derivative of J with respect to $\mathbf{W}$ to zero with $\{\mathbf{z}_n\}$ fixed gives rise to the PCA M-step (12.59).

12.18 Analysis of the number of independent parameters follows the same lines as for probabilistic PCA except that the one parameter noise covariance $\sigma^2\mathbf{I}$ is replaced by a D parameter diagonal covariance $\mathbf{\Psi}$. Thus the number of parameters is increased by $D - 1$ compared to the probabilistic PCA result (12.51) giving a total number of independent parameters of

$$D(M + 1) - M(M - 1)/2.$$

12.19 To see this we define a rotated latent space vector $\widetilde{\mathbf{z}} = \mathbf{R}\mathbf{z}$ where $\mathbf{R}$ is an $M \times M$ orthogonal matrix, and similarly defining a modified factor loading matrix $\widetilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$. Then we note that the latent space distribution $p(\mathbf{z})$ depends only on $\mathbf{z}^{\mathrm{T}}\mathbf{z} = \widetilde{\mathbf{z}}^{\mathrm{T}}\widetilde{\mathbf{z}}$, where we have used $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{I}$. Similarly, the conditional distribution of the observed variable $p(\mathbf{x}|\mathbf{z})$ depends only on $\mathbf{W}\mathbf{z} = \widetilde{\mathbf{W}}\widetilde{\mathbf{z}}$. Thus the joint distribution takes the same form for any choice of $\mathbf{R}$. This is reflected in the predictive distribution $p(\mathbf{x})$ which depends on $\mathbf{W}$ only through the quantity $\mathbf{W}\mathbf{W}^{\mathrm{T}} = \widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^{\mathrm{T}}$ and hence is also invariant to different choices of $\mathbf{R}$.

12.20 The log likelihood function is given by

$$\begin{aligned}\ln L(\boldsymbol{\mu}, \mathbf{W}, \boldsymbol{\Psi}) &= \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\boldsymbol{\mu}, \mathbf{C}) \\ &= \sum_{n=1}^{N} \left\{ -\ln |\mathbf{C}| - (\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \right\}\end{aligned}$$

where $\mathbf{C}$ is defined by (12.65). Differentiating with respect to $\boldsymbol{\mu}^{\mathrm{T}}$ and setting the derivative to zero we obtain

$$0 = \sum_{n=1}^{N} \mathbf{C}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}).$$

Pre-multiplying by $\mathbf{C}$ and re-arranging shows that $\boldsymbol{\mu}$ is given by the sample mean defined by (12.1). Taking the second derivative of the log likelihood we obtain

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\mu}^{\mathrm{T}} \partial \boldsymbol{\mu}} = -N\mathbf{C}^{-1}.$$

Since $\mathbf{C}$ is a positive definite matrix, its inverse will also be positive definite (see Appendix C) and hence the stationary point will be a unique maximum of the log likelihood.

12.21 By making use of (2.113)–(2.117) together with (12.31) and (12.64), we obtain the posterior distribution of the latent variable $\mathbf{z}$, for a given value of the observed variable $\mathbf{x}$, in the form

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{G}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}(\mathbf{x} - \overline{\mathbf{x}}).$$

where $\mathbf{G}$ is defined by (12.68). Since the data points are drawn independently from the distribution, the posterior distribution for $\mathbf{z}_n$ depends only on the observation $\mathbf{x}_n$ (for given values of the parameters). Thus (12.66) follows directly. For the second order statistic we use the general result

$$\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^{\mathrm{T}}] = \mathrm{cov}[\mathbf{z}_n] + \mathbb{E}[\mathbf{z}_n]\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}$$

from which we obtain (12.67).

12.22 **NOTE**: In PRML, Equations (12.69) and (12.70) contain minor typographical errors. On the l.h.s. $\mathbf{W}^{\mathrm{new}}$ and $\boldsymbol{\Psi}^{\mathrm{new}}$ should be $\mathbf{W}_{\mathrm{new}}$ and $\boldsymbol{\Psi}_{\mathrm{new}}$, respectively.

For the M step we first write down the complete-data log likelihood function, which takes the form

$$\begin{aligned}\ln L_{\mathrm{C}} &= \sum_{n=1}^{N} \{\ln p(\mathbf{z}_n) + \ln p(\mathbf{x}_n|\mathbf{z}_n)\} \\ &= \frac{1}{2} \sum_{n=1}^{N} \Big\{ -M \ln(2\pi) - \mathbf{z}_n^{\mathrm{T}} \mathbf{z}_n - D \ln(2\pi) - \ln |\boldsymbol{\Psi}| \\ &\qquad -(\mathbf{x}_n - \overline{\mathbf{x}} - \mathbf{W}\mathbf{z}_n)^{\mathrm{T}} \boldsymbol{\Psi}^{-1} (\mathbf{x}_n - \overline{\mathbf{x}} - \mathbf{W}\mathbf{z}_n) \Big\}.\end{aligned}$$

Now take the expectation with respect to $\{\mathbf{z}_n\}$ to give

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}\left[\ln L_{\mathrm{C}}\right] &= \frac{1}{2}\sum_{n=1}^{N}\Big\{-\ln|\boldsymbol{\Psi}| - \mathrm{Tr}\left(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{W}\right) \\ &\quad +2\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}(\mathbf{x}_n-\overline{\mathbf{x}})\Big\} - N\mathrm{Tr}\left(\mathbf{S}\boldsymbol{\Psi}^{-1}\right) + \mathrm{const.}\end{aligned}$$

where $\mathbf{S}$ is the sample covariance matrix defined by (12.3), and the constant terms are those which are independent of $\mathbf{W}$ and $\boldsymbol{\Psi}$. Recall that we are making a joint optimization with respect to $\mathbf{W}$ and $\boldsymbol{\Psi}$. Setting the derivative with respect to $\mathbf{W}^{\mathrm{T}}$ equal to zero, making use of the result (C.24), we obtain

$$0 = -2\boldsymbol{\Psi}^{-1}\mathbf{W}\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}] + 2\boldsymbol{\Psi}^{-1}\sum_{n=1}^{N}\left[(\mathbf{x}_n-\overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\right].$$

Pre-multiplying by $\boldsymbol{\Psi}$ and re-arranging we then obtain (12.69). Note that this result is independent of $\boldsymbol{\Psi}$.

Next we maximize the expected complete-data log likelihood with respect to $\boldsymbol{\Psi}$. For convenience we set the derivative with respect to $\boldsymbol{\Psi}^{-1}$ equal to zero, and make use of (C.28) to give

$$0 = N\boldsymbol{\Psi} - \mathbf{W}\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\right]\mathbf{W}^{\mathrm{T}} + 2\left[\sum_{n=1}^{N}(\mathbf{x}_n-\overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\right]\mathbf{W} - N\mathbf{S}.$$

This depends on $\mathbf{W}$, and so we can substitute for $\mathbf{W}_{\mathrm{new}}$ in the second term, using the result (12.69), which simplifies the expression. Finally, since $\boldsymbol{\Psi}$ is constrained to be diagonal, we take set all of the off-diagonal components to zero giving (12.70) as required.

12.23 The solution is given in figure 10. The model in which all parameters are shared (left) is not particularly useful, since all mixture components will have identical parameters and the resulting density model will not be any different to one offered by a single PPCA model. Different models would have arisen if only some of the parameters, e.g. the mean $\boldsymbol{\mu}$, would have been shared.

12.24 We can derive an EM algorithm by treating η in (2.160) as a latent variable. Thus given a set of i.i.d. data points, $\mathbf{X} = \{\mathbf{x}_n\}$, we define the complete-data log likelihood as

$$\ln p(\mathbf{X}, \boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \sum_{n=1}^{N}\left\{\ln\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, (\eta_n\boldsymbol{\Lambda})^{-1}\right) + \ln\mathrm{Gam}\left(\eta_n|\nu/2, \nu/2\right)\right\}$$

where $\boldsymbol{\eta}$ is an N-dimensional vector with elements η_n. The corresponding expected

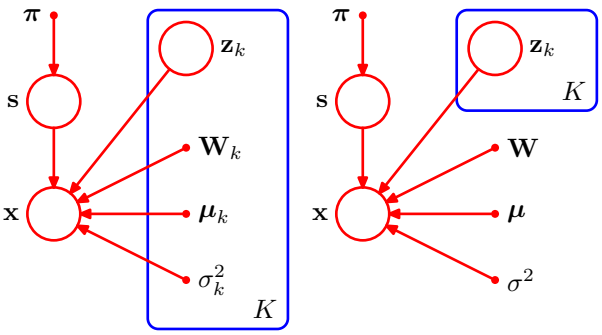


Figure 10 The left plot shows the graphical model corresponding to the general mixture of probabilistic PCA. The right plot shows the corresponding model were the parameter of all probabilist PCA models ($\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2) are shared across components. In both plots, $\mathbf{s}$ denotes the K-nomial latent variable that selects mixture components; it is governed by the parameter, $\boldsymbol{\pi}$.

complete-data log likelihood is then given by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\eta}}\left[\ln p(\mathbf{X},\boldsymbol{\eta}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)\right] = -\frac{1}{2}\sum_{n=1}^{N}\Big\{ & D(\ln(2\pi)-\mathbb{E}[\ln\eta_n]) - \ln|\boldsymbol{\Lambda}| \\ & + \mathbb{E}[\eta_n]\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu}\right) + 2\ln\Gamma(\nu/2) \\ & - \nu(\ln\nu - \ln 2) - (\nu-2)\mathbb{E}[\ln\eta_n] + \mathbb{E}[\eta_n]\Big\} \quad (316)\end{aligned}$$

where we have used results from Appendix B. In order to compute the necessary expectations, we need the distribution over $\boldsymbol{\eta}$, given by

$$\begin{aligned}p(\boldsymbol{\eta}|\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \prod_{n=1}^{N} p(\eta_n|\mathbf{x}_n,\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) \\ &\propto \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu},(\eta_n\boldsymbol{\Lambda})^{-1}\right)\mathrm{Gam}\left(\eta_n|\nu/2,\nu/2\right).\end{aligned}$$

From Section 2.3.6, we know that the factors in this product are independent Gamma distributions with parameters

$$\begin{aligned}a_n &= \frac{\nu+D}{2} \\ b_n &= \frac{\nu+(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}_n-\boldsymbol{\mu})}{2}\end{aligned}$$

and the necessary expectations are given by

$$\begin{aligned}\mathbb{E}[\eta_n] &= \frac{a_n}{b_n} \\ \mathbb{E}[\ln\eta_n] &= \psi(a_n) - \ln b_n.\end{aligned}$$

In the M step, we calculate the derivatives of (316) w.r.t. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, set these equal to

zero and solve for the respective parameter, to obtain

$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{ML}} &= \frac{\sum_{n=1}^{N}\mathbb{E}[\eta_n]\mathbf{x}_n}{\sum_{n=1}^{N}\mathbb{E}[\eta_n]} \\ \boldsymbol{\Lambda}_{\mathrm{ML}} &= \left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\eta_n](\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}\right)^{-1}\end{aligned}$$

Also for ν, we calculate the derivative of (316) and set the result equal to zero, to get

$$1+\ln\left(\frac{\nu}{2}\right)-\psi\left(\frac{\nu}{2}\right)+\frac{1}{2}\sum_{n=1}^{N}\{\mathbb{E}[\ln\eta_n]-\mathbb{E}[\eta_n]\}=0.$$

Unfortunately, there is no closed form solution w.r.t. ν, but since ν is scalar, we can afford to solve this equation numerically.

12.25 Following the discussion of section 12.2, the log likelihood function for this model can be written as

$$\begin{aligned}L(\boldsymbol{\mu},\mathbf{W},\boldsymbol{\Phi}) = &-\frac{ND}{2}\ln(2\pi)-\frac{N}{2}\ln|\mathbf{W}\mathbf{W}^{\mathrm{T}}+\boldsymbol{\Phi}| \\ &-\frac{1}{2}\sum_{n=1}^{N}\left\{(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}}+\boldsymbol{\Phi})^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right\},\end{aligned}$$

where we have used (12.43).

If we consider the log likelihood function for the transformed data set we obtain

$$\begin{aligned}L_{\mathbf{A}}(\boldsymbol{\mu},\mathbf{W},\boldsymbol{\Phi}) = &-\frac{ND}{2}\ln(2\pi)-\frac{N}{2}\ln|\mathbf{W}\mathbf{W}^{\mathrm{T}}+\boldsymbol{\Phi}| \\ &-\frac{1}{2}\sum_{n=1}^{N}\left\{(\mathbf{A}\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}}+\boldsymbol{\Phi})^{-1}(\mathbf{A}\mathbf{x}_n-\boldsymbol{\mu})\right\}.\end{aligned}$$

Solving for the maximum likelihood estimator for $\boldsymbol{\mu}$ in the usual way we obtain

$$\boldsymbol{\mu}_{\mathbf{A}}=\frac{1}{N}\sum_{n=1}^{N}\mathbf{A}\mathbf{x}_n=\mathbf{A}\overline{\mathbf{x}}=\mathbf{A}\boldsymbol{\mu}_{\mathrm{ML}}.$$

Back-substituting into the log likelihood function, and using the definition of the sample covariance matrix (12.3), we obtain

$$\begin{aligned}L_{\mathbf{A}}(\boldsymbol{\mu},\mathbf{W},\boldsymbol{\Phi}) = &-\frac{ND}{2}\ln(2\pi)-\frac{N}{2}\ln|\mathbf{W}\mathbf{W}^{\mathrm{T}}+\boldsymbol{\Phi}| \\ &-\frac{1}{2}\sum_{n=1}^{N}\mathrm{Tr}\left\{(\mathbf{W}\mathbf{W}^{\mathrm{T}}+\boldsymbol{\Phi})^{-1}\mathbf{A}\mathbf{S}\mathbf{A}^{\mathrm{T}}\right\}.\end{aligned}$$

We can cast the final term into the same form as the corresponding term in the original log likelihood function if we first define

$$\boldsymbol{\Phi}_{\mathbf{A}} = \mathbf{A}\boldsymbol{\Phi}^{-1}\mathbf{A}^{\mathrm{T}}, \qquad \mathbf{W}_{\mathbf{A}} = \mathbf{A}\mathbf{W}.$$

With these definitions the log likelihood function for the transformed data set takes the form

$$L_{\mathbf{A}}(\boldsymbol{\mu}_{\mathbf{A}}, \mathbf{W}_{\mathbf{A}}, \boldsymbol{\Phi}_{\mathbf{A}}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{W}_{\mathbf{A}}\mathbf{W}_{\mathbf{A}}^{\mathrm{T}} + \boldsymbol{\Phi}_{\mathbf{A}}| - \frac{1}{2}\sum_{n=1}^{N}\left\{(\mathbf{x}_n - \boldsymbol{\mu}_{\mathbf{A}})^{\mathrm{T}}(\mathbf{W}_{\mathbf{A}}\mathbf{W}_{\mathbf{A}}^{\mathrm{T}} + \boldsymbol{\Phi}_{\mathbf{A}})^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathbf{A}})\right\} - N\ln|\mathbf{A}|.$$

This takes the same form as the original log likelihood function apart from an additive constant $-\ln|\mathbf{A}|$. Thus the maximum likelihood solution in the new variables for the transformed data set will be identical to that in the old variables.

We now ask whether specific constraints on $\boldsymbol{\Phi}$ will be preserved by this re-scaling. In the case of probabilistic PCA the noise covariance $\boldsymbol{\Phi}$ is proportional to the unit matrix and takes the form $\sigma^2\mathbf{I}$. For this constraint to be preserved we require $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{I}$ so that $\mathbf{A}$ is an orthogonal matrix. This corresponds to a rotation of the coordinate system. For factor analysis $\boldsymbol{\Phi}$ is a diagonal matrix, and this property will be preserved if $\mathbf{A}$ is also diagonal since the product of diagonal matrices is again diagonal. This corresponds to an independent re-scaling of the coordinate system. Note that in general probabilistic PCA is not invariant under component-wise re-scaling and factor analysis is not invariant under rotation. These results are illustrated in Figure 11.

12.26 If we multiply (12.80) by $\mathbf{K}$ we obtain (12.79) so that any solution of the former will also be a solution of the latter. Let $\widetilde{\mathbf{a}}_i$ be a solution of (12.79) with eigenvalue λ_i and let $\mathbf{a}_i$ be a solution of (12.80) also having eigenvalue λ_i. If we write $\widetilde{\mathbf{a}}_i = \mathbf{a}_i + \mathbf{b}_i$ we see that $\mathbf{b}_i$ must satisfy $\mathbf{K}\mathbf{b}_i = \mathbf{0}$ and hence is an eigenvector of $\mathbf{K}$ with eigenvalue 0. It therefore satisfies

$$\sum_{n=1}^{N} b_{ni}k(\mathbf{x}_n, \mathbf{x}) = 0$$

for all values of $\mathbf{x}$. Now consider the eigenvalue projection. We see that

$$\begin{aligned}\widetilde{\mathbf{v}}_i^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) &= \sum_{n=1}^{N}\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\widetilde{a}_{ni}\boldsymbol{\phi}(\mathbf{x}_n) \\ &= \sum_{n=1}^{N}a_{ni}k(\mathbf{x}_n, \mathbf{x}) + \sum_{n=1}^{N}b_{ni}k(\mathbf{x}_n, \mathbf{x}) = \sum_{n=1}^{N}a_{ni}k(\mathbf{x}_n, \mathbf{x})\end{aligned}$$

and so both solutions give the same projections. A slightly different treatment of the relationship between (12.79) and (12.80) is given by Schölkopf (1998).

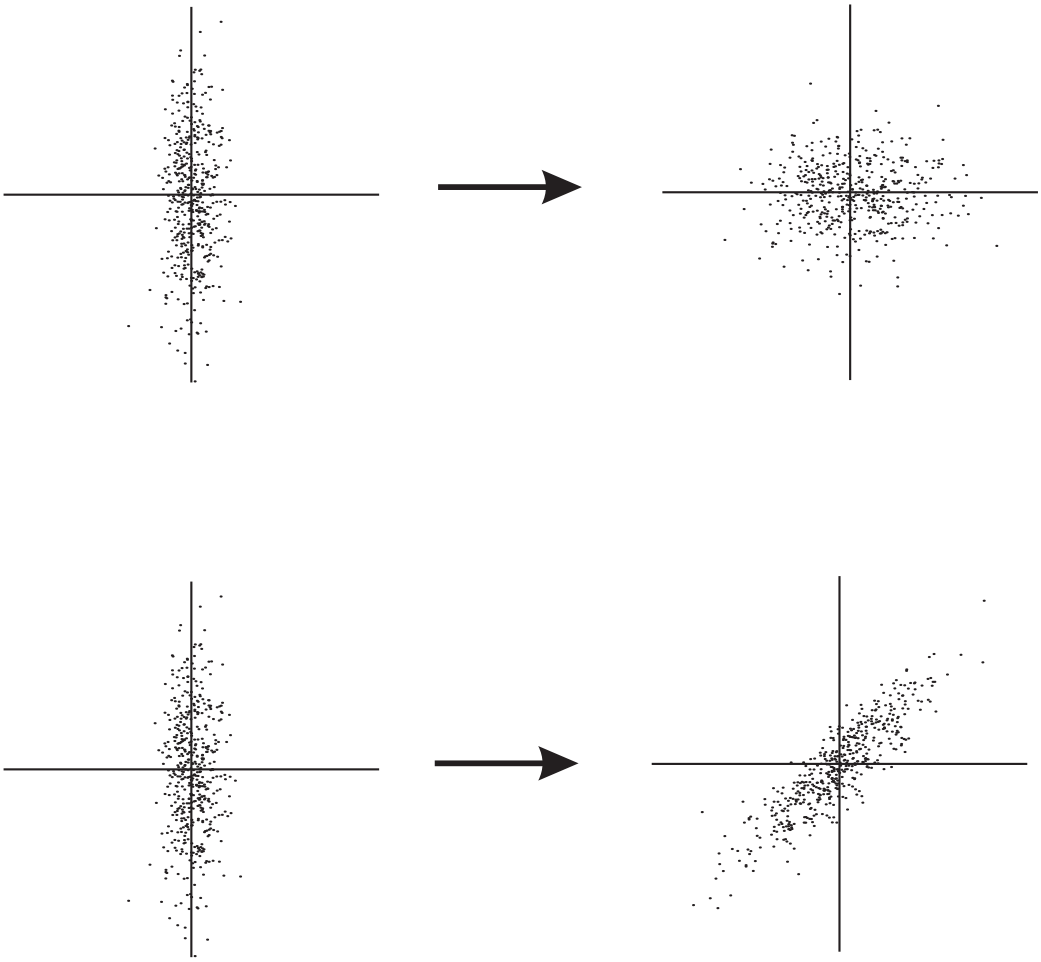


Figure 11 Factor analysis is covariant under a componentwise re-scaling of the data variables (top plots), while PCA and probabilistic PCA are covariant under rotations of the data space coordinates (lower plots).

12.27 In the case of the linear kernel, we can rewrite the l.h.s. of (12.80) as

$$\begin{aligned}\mathbf{K}\mathbf{a}_i &= \sum_{m=1}^{N} k(\mathbf{x}_n, \mathbf{x}_m) a_{im} \\ &= \sum_{m=1}^{N} \mathbf{x}_n^{\mathrm{T}} \mathbf{x}_m a_{mi}\end{aligned}$$

and substituting this in (12.80), we get

$$\sum_{m=1}^{N} \mathbf{x}_n^{\mathrm{T}} \mathbf{x}_m a_{mi} = \lambda_i N \mathbf{a}_i.$$

Next, we left-multiply both sides by $\mathbf{x}_n$ and sum over n to obtain

$$N\mathbf{S} \sum_{m=1}^{N} \mathbf{x}_m a_{im} = \lambda_i N \sum_{n=1}^{N} \mathbf{x}_n a_{in}.$$

Finally, we divide both sides by N and define

$$\mathbf{u}_i = \sum_{n=1}^{N} \mathbf{x}_n a_{in}$$

to recover (12.17).

12.28 If we assume that the function $y = f(x)$ is *strictly* monotonic, which is necessary to exclude the possibility for spikes of infinite density in $p(y)$, we are guaranteed that the inverse function $x = f^{-1}(y)$ exists. We can then use (1.27) to write

$$p(y) = q(f^{-1}(y)) \left| \frac{\mathrm{d}f^{-1}}{\mathrm{d}y} \right|. \tag{317}$$

Since the only restriction on f is that it is monotonic, it can distribute the probability mass over x arbitrarily over y. This is illustrated in Figure 1 on page 9, as a part of Solution 1.4. From (317) we see directly that

$$|f'(x)| = \frac{q(x)}{p(f(x))}.$$

12.29 **NOTE**: In the 1^{st} printing of PRML, this exercise contains two mistakes. In the second half of the exercise, we require that y_1 is symmetrically distributed around 0, not just that $-1 \leqslant y_1 \leqslant 1$. Moreover, $y_2 = y_1^2$ (not $y_2 = y_2^2$).

If z_1 and z_2 are independent, then

$$\begin{aligned}\mathrm{cov}[z_1, z_2] &= \iint (z_1 - \bar{z}_1)(z_2 - \bar{z}_2)p(z_1, z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2 \\ &= \iint (z_1 - \bar{z}_1)(z_2 - \bar{z}_2)p(z_1)p(z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2 \\ &= \int (z_1 - \bar{z}_1)p(z_1)\,\mathrm{d}z_1 \int (z_2 - \bar{z}_2)p(z_2)\,\mathrm{d}z_2 \\ &= 0,\end{aligned}$$

where

$$\bar{z}_i = \mathbb{E}[z_i] = \int z_i p(z_i)\,\mathrm{d}z_i.$$

For y_2 we have

$$p(y_2|y_1) = \delta(y_2 - y_1^2),$$

i.e., a spike of probability mass one at y_1^2, which is clearly dependent on y_1. With $\bar{y}_i$ defined analogously to $\bar{z}_i$ above, we get

$$\begin{aligned}\mathrm{cov}[y_1, y_2] &= \iint (y_1 - \bar{y}_1)(y_2 - \bar{y}_2)p(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &= \iint y_1(y_2 - \bar{y}_2)p(y_2|y_1)p(y_1)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &= \int (y_1^3 - y_1\bar{y}_2)p(y_1)\,\mathrm{d}y_1 \\ &= 0,\end{aligned}$$

where we have used the fact that all odd moments of y_1 will be zero, since it is symmetric around zero.

Chapter 13 Sequential Data

13.1 Since the arrows on the path from x_m to x_n, with $m < n - 1$, will meet head-to-tail at x_{n-1}, which is in the conditioning set, all such paths are blocked by x_{n-1} and hence (13.3) holds.

The same argument applies in the case depicted in Figure 13.4, with the modification that $m < n - 2$ and that paths are blocked by x_{n-1} or x_{n-2}.

13.2 We first of all find the joint distribution $p(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ by marginalizing over the variables $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N$, to give

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_n) &= \sum_{\mathbf{x}_{n+1}} \cdots \sum_{\mathbf{x}_N} p(\mathbf{x}_1, \ldots, \mathbf{x}_N) \\ &= \sum_{\mathbf{x}_{n+1}} \cdots \sum_{\mathbf{x}_N} p(\mathbf{x}_1) \prod_{m=2}^{N} p(\mathbf{x}_m|\mathbf{x}_{m-1}) \\ &= p(\mathbf{x}_1) \prod_{m=2}^{n} p(\mathbf{x}_m|\mathbf{x}_{m-1}). \end{aligned}$$

Now we evaluate the required conditional distribution

$$\begin{aligned} p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) &= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_n)}{\sum_{\mathbf{x}_n} p(\mathbf{x}_1, \ldots, \mathbf{x}_n)} \\ &= \frac{p(\mathbf{x}_1) \prod_{m=2}^{n} p(\mathbf{x}_m|\mathbf{x}_{m-1})}{\sum_{\mathbf{x}_n} p(\mathbf{x}_1) \prod_{m=2}^{n} p(\mathbf{x}_m|\mathbf{x}_{m-1})}. \end{aligned}$$

We now note that any factors which do not depend on $\mathbf{x}_n$ will cancel between numerator and denominator, giving

$$\begin{aligned} p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) &= \frac{p(\mathbf{x}_n|\mathbf{x}_{n-1})}{\sum_{\mathbf{x}_n} p(\mathbf{x}_n|\mathbf{x}_{n-1})} \\ &= p(\mathbf{x}_n|\mathbf{x}_{n-1}) \end{aligned}$$

as required.

For the second order Markov model, the joint distribution is given by (13.4). The marginal distribution over the variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is given by

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_n) &= \sum_{\mathbf{x}_{n+1}} \cdots \sum_{\mathbf{x}_N} p(\mathbf{x}_1, \ldots, \mathbf{x}_N) \\ &= \sum_{\mathbf{x}_{n+1}} \cdots \sum_{\mathbf{x}_N} p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1) \prod_{m=3}^{N} p(\mathbf{x}_m|\mathbf{x}_{m-1}, \mathbf{x}_{m-2}) \\ &= p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1) \prod_{m=3}^{n} p(\mathbf{x}_m|\mathbf{x}_{m-1}, \mathbf{x}_{m-2}). \end{aligned}$$

The required conditional distribution is then given by

$$p(\mathbf{x}_n|\mathbf{x}_1,\ldots,\mathbf{x}_{n-1}) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_n)}{\sum_{\mathbf{x}_n} p(\mathbf{x}_1,\ldots,\mathbf{x}_n)}$$
$$= \frac{p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)\prod_{m=3}^{n} p(\mathbf{x}_m|\mathbf{x}_{m-1},\mathbf{x}_{m-2})}{\sum_{\mathbf{x}_n} p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)\prod_{m=3}^{n} p(\mathbf{x}_m|\mathbf{x}_{m-1},\mathbf{x}_{m-2})}.$$

Again, cancelling factors independent of $\mathbf{x}_n$ between numerator and denominator we obtain

$$p(\mathbf{x}_n|\mathbf{x}_1,\ldots,\mathbf{x}_{n-1}) = \frac{p(\mathbf{x}_n|\mathbf{x}_{n-1},\mathbf{x}_{n-2})}{\sum_{\mathbf{x}_n} p(\mathbf{x}_n|\mathbf{x}_{n-1},\mathbf{x}_{n-2})}$$
$$= p(\mathbf{x}_n|\mathbf{x}_{n-1},\mathbf{x}_{n-2}).$$

Thus the prediction at step n depends only on the observations at the two previous steps $\mathbf{x}_{n-1}$ and $\mathbf{x}_{n-2}$ as expected.

13.3 From Figure 13.5 we see that for any two variables $\mathbf{x}_n$ and $\mathbf{x}_m$, $m \neq n$, there is a path between the corresponding nodes that will only pass through one or more nodes corresponding to $\mathbf{z}$ variables. None of these nodes will be in the conditioning set and the edges on the path meet head-to-tail. Thus, there will be an unblocked path between $\mathbf{x}_n$ and $\mathbf{x}_m$ and the model will not satisfy any conditional independence or finite order Markov properties.

13.4 The learning of $\mathbf{w}$ would follow the scheme for maximum learning described in Section 13.2.1, with $\mathbf{w}$ replacing $\boldsymbol{\phi}$. As discussed towards the end of Section 13.2.1, the precise update formulae would depend on the form of regression model used and how it is being used.

The most obvious situation where this would occur is in a HMM similar to that depicted in Figure 13.18, where the emmission densities not only depends on the latent variable $\mathbf{z}$, but also on some input variable $\mathbf{u}$. The regression model could then be used to map $\mathbf{u}$ to $\mathbf{x}$, depending on the state of the latent variable $\mathbf{z}$.

Note that when a nonlinear regression model, such as a neural network, is used, the M-step for $\mathbf{w}$ may not have closed form.

13.5 Consider first the maximization with respects to the components π_k of $\boldsymbol{\pi}$. To do this we must take account of the summation constraint

$$\sum_{k=1}^{K} \pi_k = 1.$$

We therefore first omit terms from $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}})$ which are independent of $\boldsymbol{\pi}$, and then add a Lagrange multiplier term to enforce the constraint, giving the following function to be maximized

$$\widetilde{Q} = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right).$$

Setting the derivative with respect to π_k equal to zero we obtain

$$0 = \gamma(z_{1k}) \frac{1}{\pi_k} + \lambda. \tag{318}$$

We now multiply through by π_k and then sum over k and make use of the summation constraint to give

$$\lambda = -\sum_{k=1}^{K} \gamma(z_{1k}).$$

Substituting back into (318) and solving for λ we obtain (13.18).

For the maximization with respect to $\mathbf{A}$ we follow the same steps and first omit terms from $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}})$ which are independent of $\mathbf{A}$, and then add appropriate Lagrange multiplier terms to enforce the summation constraints. In this case there are K constraints to be satisfied since we must have

$$\sum_{k=1}^{K} A_{jk} = 1$$

for $j = 1, \ldots, K$. We introduce K Lagrange multipliers λ_j for $j = 1, \ldots, K$, and maximize the following function

$$\widehat{Q} = \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} + \sum_{j=1}^{K} \lambda_j \left(\sum_{k=1}^{K} A_{jk} - 1 \right).$$

Setting the derivative of $\widehat{Q}$ with respect to A_{jk} to zero we obtain

$$0 = \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk}) \frac{1}{A_{jk}} + \lambda_j. \tag{319}$$

Again we multiply through by A_{jk} and then sum over k and make use of the summation constraint to give

$$\lambda_j = -\sum_{n=2}^{N} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}).$$

Substituting for λ_j in (319) and solving for A_{jk} we obtain (13.19).

13.6 Suppose that a particular element π_k of $\boldsymbol{\pi}$ has been initialized to zero. In the first E-step the quantity $\alpha(z_{1k})$ is given from (13.37) by

$$\alpha(z_{1k}) = \pi_k p(\mathbf{x}_1|\boldsymbol{\phi}_k)$$

and so will be zero. From (13.33) we see that $\gamma(z_{1k})$ will also be zero, and hence in the next M-step the new value of π_k, given by (13.18) will again be zero. Since this is true for any subsequent EM cycle, this quantity will remain zero throughout.

Similarly, suppose that an element A_{jk} of $\mathbf{A}$ has been set initially to zero. From (13.43) we see that $\xi(z_{n-1,j}, z_{nk})$ will be zero since $p(z_{nk}|z_{n-1,j}) \equiv A_{jk}$ equals zero. In the subsequent M-step, the new value of A_{jk} is given by (13.19) and hence will also be zero.

13.7 Using the expression (13.17) for $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}})$ we see that the parameters of the Gaussian emission densities appear only in the last term, which takes the form

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\ln p(\mathbf{x}_n|\boldsymbol{\phi}_k) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_j)$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right\}.$$

We now maximize this quantity with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. Setting the derivative with respect to $\boldsymbol{\mu}_k$ to zero and re-arranging we obtain (13.20). Next if we define

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$
$$\widehat{\mathbf{S}}_k = \sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

then we can rewrite the final term from $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}})$ in the form

$$-\frac{N_k D}{2}\ln(2\pi) - \frac{N_k}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_k^{-1}\widehat{\mathbf{S}}_k\right).$$

Differentiating this w.r.t. $\boldsymbol{\Sigma}_k^{-1}$, using results from Appendix C, we obtain (13.21).

13.8 Only the final term of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}$ given by (13.17) depends on the parameters of the emission model. For the multinomial variable $\mathbf{x}$, whose D components are all zero except for a single entry of 1,

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\ln p(\mathbf{x}_n|\boldsymbol{\phi}_k) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\sum_{i=1}^{D}x_{ni}\ln\mu_{ki}.$$

Now when we maximize with respect to μ_{ki} we have to take account of the constraints that, for each value of k the components of μ_{ki} must sum to one. We therefore introduce Lagrange multipliers $\{\lambda_k\}$ and maximize the modified function given by

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\sum_{i=1}^{D}x_{ni}\ln\mu_{ki}+\sum_{k=1}^{K}\lambda_k\left(\sum_{i=1}^{D}\mu_{ki}-1\right).$$

Setting the derivative with respect to μ_{ki} to zero we obtain

$$0=\sum_{n=1}^{N}\gamma(z_{nk})\frac{x_{ni}}{\mu_{ki}}+\lambda_k.$$

Multiplying through by μ_{ki}, summing over i, and making use of the constraint on μ_{ki} together with the result $\sum_i x_{ni}=1$ we have

$$\lambda_k=-\sum_{n=1}^{N}\gamma(z_{nk}).$$

Finally, back-substituting for λ_k and solving for μ_{ki} we again obtain (13.23).

Similarly, for the case of a multivariate Bernoulli observed variable $\mathbf{x}$ whose D components independently take the value 0 or 1, using the standard expression for the multivariate Bernoulli distribution we have

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\ln p(\mathbf{x}_n|\boldsymbol{\phi}_k)$$
$$=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\sum_{i=1}^{D}\left\{x_{ni}\ln\mu_{ki}+(1-x_{ni})\ln(1-\mu_{ki})\right\}.$$

Maximizing with respect to μ_{ki} we obtain

$$\mu_{ki}=\frac{\displaystyle\sum_{n=1}^{N}\gamma(z_{nk})x_{ni}}{\displaystyle\sum_{n=1}^{N}\gamma(z_{nk})}$$

which is equivalent to (13.23).

13.9 We can verify all these independence properties using d-separation by refering to Figure 13.5.

(13.24) follows from the fact that arrows on paths from any of $\mathbf{x}_1,\ldots,\mathbf{x}_n$ to any of $\mathbf{x}_{n+1},\ldots,\mathbf{x}_N$ meet head-to-tail or tail-to-tail at $\mathbf{z}_n$, which is in the conditioning set.

(13.25) follows from the fact that arrows on paths from any of $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ to $\mathbf{x}_n$ meet head-to-tail at $\mathbf{z}_n$, which is in the conditioning set.

(13.26) follows from the fact that arrows on paths from any of $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ to $\mathbf{z}_n$ meet head-to-tail or tail-to-tail at $\mathbf{z}_{n-1}$, which is in the conditioning set.

(13.27) follows from the fact that arrows on paths from $\mathbf{z}_n$ to any of $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N$ meet head-to-tail at $\mathbf{z}_{n+1}$, which is in the conditioning set.

(13.28) follows from the fact that arrows on paths from $\mathbf{x}_{n+1}$ to any of $\mathbf{x}_{n+2}, \ldots, \mathbf{x}_N$ to meet tail-to-tail at $\mathbf{z}_{n+1}$, which is in the conditioning set.

(13.29) follows from (13.24) and the fact that arrows on paths from any of $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ to $\mathbf{x}_n$ meet head-to-tail or tail-to-tail at $\mathbf{z}_{n-1}$, which is in the conditioning set.

(13.30) follows from the fact that arrows on paths from any of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ to $\mathbf{x}_{N+1}$ meet head-to-tail at $\mathbf{z}_{N+1}$, which is in the conditioning set.

(13.31) follows from the fact that arrows on paths from any of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ to $\mathbf{z}_{N+1}$ meet head-to-tail or tail-to-tail at $\mathbf{z}_N$, which is in the conditioning set.

13.10 We begin with the expression (13.10) for the joint distribution of observed and latent variables in the hidden Markov model, reproduced here for convenience

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1)\left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1})\right]\prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m)$$

where we have omitted the parameters in order to keep the notation uncluttered. By marginalizing over all of the latent variables except $\mathbf{z}_n$ we obtain the joint distribution of the remaining variables, which can be factorized in the form

$$\begin{aligned} p(\mathbf{X}, \mathbf{z}_n) &= \sum_{\mathbf{z}_1}\cdots\sum_{\mathbf{z}_{n-1}}\sum_{\mathbf{z}_{n+1}}\cdots\sum_{\mathbf{z}_N} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \\ &= \left[\sum_{\mathbf{z}_1}\cdots\sum_{\mathbf{z}_{n-1}} p(\mathbf{z}_1)\prod_{m=2}^{n} p(\mathbf{z}_m|\mathbf{z}_{m-1})\prod_{l=1}^{n} p(\mathbf{x}_l|\mathbf{z}_l)\right] \\ &\quad \times \left[\sum_{\mathbf{z}_{n+1}}\cdots\sum_{\mathbf{z}_N}\prod_{m=n+1}^{N} p(\mathbf{z}_m|\mathbf{z}_{m-1})\prod_{l=n+1}^{N} p(\mathbf{x}_l|\mathbf{z}_l)\right]. \end{aligned}$$

The first factor in square brackets on the r.h.s. we recognize as $p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)$. Next we note from the product rule that

$$p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n) = \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_n)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)}.$$

Thus we can identify the second term in square brackets with the conditional distribution $p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)$. However, we note that the second term in square brackets does not depend on $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Thus we have the result

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n).$$

Dividing both sides by $p(\mathbf{z}_n)$ we obtain

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n).$$

which is the required result (13.24).

Similarly, from (13.10) we have

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)$$

It follows that

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-1}|\mathbf{x}_n, \mathbf{z}_n) &= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n)}{p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n)} \\ &= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})}{p(\mathbf{z}_n)} \end{aligned}$$

where we see that the right hand side is independent of $\mathbf{x}_n$, and hence the left hand side must be also. We therefore have the following conditional independence property

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-1}|\mathbf{x}_n, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-1}|\mathbf{z}_n).$$

Marginalizing both sides of this result over $\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}$ then gives the required result (13.25).

Again, from the joint distribution we can write

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1}).$$

We therefore have

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-2}|\mathbf{z}_{n-1}, \mathbf{z}_n) &= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_n)}{p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{z}_{n-1})} \\ &= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-1})}{p(\mathbf{z}_{n-1})} \end{aligned}$$

where we see that the right hand side is independent of $\mathbf{z}_n$ and hence the left hand side must be also. This implies the conditional independence property

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-2}|\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_1, \ldots, \mathbf{z}_{n-2}|\mathbf{z}_{n-1}).$$

Marginalizing both sides with respect to $\mathbf{z}_1, \ldots, \mathbf{z}_{n-2}$ then gives the required result (13.26).

To prove (13.27) we marginalize the both sides of the expression (13.10) for the joint distribution with respect to the variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to give

$$\begin{aligned} p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N, \mathbf{z}_n, \mathbf{z}_{n+1}) = &\left[\sum_{\mathbf{z}_1} \cdots \sum_{\mathbf{z}_{n-1}} p(\mathbf{z}_1) \prod_{m=2}^{n} p(\mathbf{z}_m|\mathbf{z}_{m-1}) \prod_{l=1}^{n} p(\mathbf{x}_l|\mathbf{z}_l)\right] \\ &p(\mathbf{z}_{n+1}|\mathbf{z}_n)p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) \left[\sum_{\mathbf{z}_{n+1}} \cdots \sum_{\mathbf{z}_N} \prod_{m=n+2}^{N} p(\mathbf{z}_m|\mathbf{z}_{m-1}) \prod_{l=n+1}^{N} p(\mathbf{x}_l|\mathbf{z}_l)\right]. \end{aligned}$$

The first factor in square brackets is just

$$\sum_{\mathbf{z}_1} \cdots \sum_{\mathbf{z}_{n-1}} p(\mathbf{z}_1, \ldots, \mathbf{z}_n) = p(\mathbf{z}_n)$$

and so by the product rule the final factor in square brackets must be

$$p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n, \mathbf{z}_{n+1}).$$

However, the second factor in square brackets is itself independent of $\mathbf{z}_n$ which proves (13.27).

To prove (13.28) we first note that the decomposition of the joint distribution $p(\mathbf{X}, \mathbf{Z})$ implies the following factorization

$$p(\mathbf{X}, \mathbf{Z}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n) p(\mathbf{z}_{n+1}|\mathbf{z}_n) p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) \\ p(\mathbf{x}_{n+2}, \ldots, \mathbf{x}_N, \mathbf{z}_{n+1}, \ldots, \mathbf{z}_N | \mathbf{z}_{n+1}).$$

Next we make use of

$$\begin{aligned} p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N, \mathbf{z}_{n+1}) &= \sum_{\mathbf{x}_1} \cdots \sum_{\mathbf{x}_n} \sum_{\mathbf{z}_1} \cdots \sum_{\mathbf{z}_n} \sum_{\mathbf{z}_{n+2}} \cdots \sum_{\mathbf{z}_N} p(\mathbf{X}, \mathbf{Z}) \\ &= p(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) p(\mathbf{x}_{n+2}, \ldots, \mathbf{x}_N | \mathbf{z}_{n+1}). \end{aligned}$$

If we now divide both sides by $p(\mathbf{x}_{n+1}, \mathbf{z}_{n+1})$ we obtain

$$p(\mathbf{x}_{n+2}, \ldots, \mathbf{x}_N | \mathbf{z}_{n+1}, \mathbf{x}_{n+1}) = p(\mathbf{x}_{n+2}, \ldots, \mathbf{x}_N | \mathbf{z}_{n+1})$$

as required.

To prove (13.30) we first use the expression for the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ to give

$$\begin{aligned} p(\mathbf{x}_{N+1}, \mathbf{X}, \mathbf{z}_{N+1}) &= \sum_{\mathbf{z}_1} \cdots \sum_{\mathbf{z}_N} p(\mathbf{X}, \mathbf{Z}) p(\mathbf{z}_{N+1}|\mathbf{z}_N) p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \\ &= p(\mathbf{X}, \mathbf{z}_{N+1}) p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \end{aligned}$$

from which it follows that

$$p(\mathbf{x}_{N+1}|\mathbf{X}, \mathbf{z}_{N+1}) = p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1})$$

as required.

To prove (13.31) we first use the expression for the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ to give

$$\begin{aligned} p(\mathbf{z}_{N+1}, \mathbf{X}, \mathbf{z}_N) &= \sum_{\mathbf{z}_1} \cdots \sum_{\mathbf{z}_{N-1}} p(\mathbf{X}, \mathbf{Z}) p(\mathbf{z}_{N+1}|\mathbf{z}_N) \\ &= p(\mathbf{X}, \mathbf{z}_N) p(\mathbf{z}_{N+1}|\mathbf{z}_N) \end{aligned}$$

from which it follows that

$$p(\mathbf{z}_{N+1}|\mathbf{z}_N, \mathbf{X}) = p(\mathbf{z}_{N+1}|\mathbf{z}_N)$$

as required.

Finally, to prove (13.29) we first marginalize both sides of the joint distribution (13.10) with respect to $\mathbf{z}_1, \ldots, \mathbf{z}_{n-2}, \mathbf{z}_{n+1}, \ldots \mathbf{z}_N$ to give

$$\begin{aligned} p(\mathbf{X}, \mathbf{z}_{n-1}, \mathbf{z}_n) &= \left[\sum_{\mathbf{z}_1} \cdots \sum_{\mathbf{z}_{n-2}} p(\mathbf{z}_1) \prod_{m=2}^{n-1} p(\mathbf{z}_m|\mathbf{z}_{m-1}) \prod_{l=1}^{n-1} p(\mathbf{x}_l|\mathbf{z}_l)\right] \\ &\quad p(\mathbf{z}_n|\mathbf{z}_{n-1}) p(\mathbf{x}_n|\mathbf{z}_n) \\ &\quad \left[\sum_{\mathbf{z}_{n+1}} \cdots \sum_{\mathbf{z}_N} \prod_{m=n+1}^{N} p(\mathbf{z}_m|\mathbf{z}_{m-1}) \prod_{l=n+1}^{N} p(\mathbf{x}_l|\mathbf{z}_l)\right]. \end{aligned}$$

The first factor in square brackets is $p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_{n-1})$. The second factor in square brackets is

$$\sum_{\mathbf{z}_{n+1}} \cdots \sum_{\mathbf{z}_N} p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N, \mathbf{z}_n, \ldots, \mathbf{z}_N) = p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N, \mathbf{z}_n).$$

Thus we have

$$\begin{aligned} p(\mathbf{X}, \mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_{n-1}) \\ p(\mathbf{z}_n|\mathbf{z}_{n-1}) p(\mathbf{x}_n|\mathbf{z}_n) p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N, \mathbf{z}_n) \end{aligned}$$

and dividing both sides by $p(\mathbf{z}_n, \mathbf{z}_{n-1}) = p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{z}_{n-1})$ we obtain (13.28).

The final conclusion from all of this exhausting algebra is that it is much easier simply to draw a graph and apply the d-separation criterion!

13.11 From the first line of (13.43), we have

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}).$$

This corresponds to the distribution over the variables associated with factor f_n in Figure 13.5, i.e. $\mathbf{z}_{n-1}$ and $\mathbf{z}_n$.

From (8.69), (8.72), (13.50) and (13.52), we have

$$\begin{aligned} p(\mathbf{z}_{n-1}, \mathbf{z}_n) &\propto f_n(\mathbf{z}_{n-1}, \mathbf{z}_n, \mathbf{x}_n) \mu_{\mathbf{z}_{n-1} \to f_n}(\mathbf{z}_{n-1}) \mu_{\mathbf{z}_n \to fn}(\mathbf{z}_n) \\ &= p(\mathbf{z}_n|\mathbf{z}_{n-1}) p(\mathbf{x}_n|\mathbf{z}_n) \mu_{f_{n-1} \to \mathbf{z}_{n-1}}(\mathbf{z}_{n-1}) \mu_{f_{n+1} \to \mathbf{z}_n}(\mathbf{z}_n) \\ &= p(\mathbf{z}_n|\mathbf{z}_{n-1}) p(\mathbf{x}_n|\mathbf{z}_n) \alpha(\mathbf{z}_{n-1}) \beta(\mathbf{z}_n). \end{aligned} \tag{320}$$

In order to normalize this, we use (13.36) and (13.41) to obtain

$$\begin{aligned} \sum_{\mathbf{z}_n} \sum_{\mathbf{z}_{n-1}} p(\mathbf{z}_{n-1}, \mathbf{z}_n) &= \sum_{\mathbf{z}_n} \beta(\mathbf{z}_n) p(\mathbf{x}_n|\mathbf{z}_n \sum_{\mathbf{z}_{n-1}} p(\mathbf{z}_n|\mathbf{z}_{n-1}) \alpha(\mathbf{z}_n) \\ &= \sum_{\mathbf{z}_n} \beta(\mathbf{z}_n) \alpha(\mathbf{z}_n) = p(\mathbf{X}) \end{aligned}$$

which together with (320) give (13.43).

13.12 First of all, note that for every observed variable there is a corresponding latent variable, and so for every sequence $\mathbf{X}^{(r)}$ of observed variables there is a corresponding sequence $\mathbf{Z}^{(r)}$ of latent variables. The sequences are assumed to be independent given the model parameters, and so the joint distribution of all latent and observed variables will be given by

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \prod_{r=1}^{R} p(\mathbf{X}^{(r)}, \mathbf{Z}^{(r)}|\boldsymbol{\theta})$$

where $\mathbf{X}$ denotes $\{\mathbf{X}^{(r)}\}$ and $\mathbf{Z}$ denotes $\{\mathbf{Z}^{(r)}\}$. Using the sum and product rules of probability we then see that posterior distribution for the latent sequences then factorizes with respect to those sequences, so that

$$\begin{aligned}
p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) &= \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})} \\
&= \frac{\prod_{r=1}^{R} p(\mathbf{X}^{(r)}, \mathbf{Z}^{(r)}|\boldsymbol{\theta})}{\sum_{\mathbf{Z}^{(1)}} \cdots \sum_{\mathbf{Z}^{(R)}} \prod_{r=1}^{R} p(\mathbf{X}^{(r)}, \mathbf{Z}^{(r)}|\boldsymbol{\theta})} \\
&= \prod_{r=1}^{R} \left\{ \frac{p(\mathbf{X}^{(r)}, \mathbf{Z}^{(r)}|\boldsymbol{\theta})}{\sum_{\mathbf{Z}^{(r)}} p(\mathbf{X}^{(r)}, \mathbf{Z}^{(r)}|\boldsymbol{\theta})} \right\} \\
&= \prod_{r=1}^{R} p(\mathbf{Z}^{(r)}|\mathbf{X}^{(r)}, \boldsymbol{\theta}).
\end{aligned}$$

Thus the evaluation of the posterior distribution of the latent variables, corresponding to the E-step of the EM algorithm, can be done independently for each of the sequences (using the standard alpha-beta recursions).

Now consider the M-step. We use the posterior distribution computed in the E-step using $\boldsymbol{\theta}_{\text{old}}$ to evaluate the expectation of the complete-data log likelihood. From our

expression for the joint distribution we see that this is given by

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}}) &= \mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\right] \\ &= \mathbb{E}_{\mathbf{Z}}\left[\sum_{r=1}^{R} \ln p(\mathbf{X}^{(r)}, \mathbf{Z}^{(r)}|\boldsymbol{\theta})\right] \\ &= \sum_{r=1}^{R} p(\mathbf{Z}^{(r)}|\mathbf{X}^{(r)}, \boldsymbol{\theta}_{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \\ &= \sum_{r=1}^{R}\sum_{k=1}^{K} \gamma(z_{1k}^{(r)}) \ln \pi_k + \sum_{r=1}^{R}\sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} \xi(z_{n-1,j}^{(r)}, z_{n,k}^{(r)}) \ln A_{jk} \\ &\quad + \sum_{r=1}^{R}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}^{(r)}) \ln p(\mathbf{x}_n^{(r)}|\boldsymbol{\phi}_k). \end{aligned}$$

We now maximize this quantity with respect to $\boldsymbol{\pi}$ and $\mathbf{A}$ in the usual way, with Lagrange multipliers to take account of the summation constraints (see Solution 13.5), yielding (13.124) and (13.125). The M-step results for the mean of the Gaussian follow in the usual way also (see Solution 13.7).

13.13 Using (8.64), we can rewrite (13.50) as

$$\alpha(\mathbf{z}_n) = \sum_{\mathbf{z}_1,\ldots,\mathbf{z}_{n-1}} F_n(\mathbf{z}_n, \{\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}\}), \tag{321}$$

where $F_n(\cdot)$ is the product of all factors connected to $\mathbf{z}_n$ via f_n, including f_n itself (see Figure 13.15), so that

$$F_n(\mathbf{z}_n, \{\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}\}) = h(\mathbf{z}_1) \prod_{i=2}^{n} f_i(\mathbf{z}_i, \mathbf{z}_{i-1}), \tag{322}$$

where we have introduced $h(\mathbf{z}_1)$ and $f_i(\mathbf{z}_i, \mathbf{z}_{i-1})$ from (13.45) and (13.46), respectively. Using the corresponding r.h.s. definitions and repeatedly applying the product rule, we can rewrite (322) as

$$F_n(\mathbf{z}_n, \{\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}\}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n).$$

Applying the sum rule, summing over $\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}$ as on the r.h.s. of (321), we obtain (13.34).

13.14 **NOTE**: In PRML, the reference to (8.67) should refer to (8.64).

This solution largely follows Solution 13.13. Using (8.64), we can rewrite (13.52) as

$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1},\ldots,\mathbf{z}_N} F_{n+1}(\mathbf{z}_n, \{\mathbf{z}_{n+1}, \ldots, \mathbf{z}_N\}), \tag{323}$$

where $F_{n+1}(\cdot)$ is the product of all factors connected to $\mathbf{z}_n$ via f_{n+1}, including f_{n+1} itself so that

$$\begin{aligned} F_{n+1}(\mathbf{z}_n, \{\mathbf{z}_{n+1}, \ldots, \mathbf{z}_N\}) &= \prod_{i=n+1}^{N} f_i(\mathbf{z}_{i-1}, \mathbf{z}_i) \\ &= p(\mathbf{z}_{n+1}|\mathbf{z}_n)p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) \prod_{i=n+2}^{N} f_i(\mathbf{z}_{i-1}, \mathbf{z}_i) \\ &= p(\mathbf{z}_{n+1}|\mathbf{z}_n)p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) \cdots p(\mathbf{z}_N|\mathbf{z}_{N-1})p(\mathbf{x}_N|\mathbf{z}_N) \end{aligned} \tag{324}$$

where we have used (13.46). Repeatedly applying the product rule, we can rewrite (324) as

$$F_{n+1}(\mathbf{z}_n, \{\mathbf{z}_{n+1}, \ldots, \mathbf{z}_N\}) = p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N, \mathbf{z}_{n+1}, \ldots, \mathbf{z}_N|\mathbf{z}_n).$$

Substituting this into (323) and summing over $\mathbf{z}_{n+1}, \ldots, \mathbf{z}_N$, we obtain (13.35).

13.15 **NOTE**: In the 1st printing of PRML, there are typographic errors in (13.65); c_n should be c_n^{-1} and $p(\mathbf{z}_n|\mathbf{z}_{-1})$ should be $p(\mathbf{z}_n|\mathbf{z}_{n-1})$ on the r.h.s.

We can use (13.58), (13.60) and (13.63) to rewrite (13.33) as

$$\begin{aligned} \gamma(\mathbf{z}_n) &= \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})} \\ &= \frac{\widehat{\alpha}(\mathbf{z}_n)\left(\prod_{m=1}^{n} c_m\right)\left(\prod_{l=n+1}^{N} c_l\right)\widehat{\beta}(\mathbf{z}_n)}{p(\mathbf{X})} \\ &= \frac{\widehat{\alpha}(\mathbf{z}_n)\left(\prod_{m=1}^{N} c_m\right)\widehat{\beta}(\mathbf{z}_n)}{p(\mathbf{X})} \\ &= \widehat{\alpha}(\mathbf{z}_n)\widehat{\beta}(\mathbf{z}_n). \end{aligned}$$

We can rewrite (13.43) in a similar fashion:

$$\begin{aligned} \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= \frac{\alpha(\mathbf{z}_n)p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})\beta(\mathbf{z}_n)}{p(\mathbf{X})} \\ &= \frac{\widehat{\alpha}(\mathbf{z}_n)\left(\prod_{m=1}^{n-1} c_m\right)\left(\prod_{l=n+1}^{N} c_l\right)\widehat{\beta}(\mathbf{z}_n)p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})}{p(\mathbf{X})} \\ &= c_n^{-1}\widehat{\alpha}(\mathbf{z}_n)p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})\widehat{\beta}(\mathbf{z}_n). \end{aligned}$$

13.16 **NOTE**: In the 1st printing of PRML, $\ln p(\mathbf{x}_{+1}|\mathbf{z}_n)$ should be $\ln p(\mathbf{z}_{n+1}|\mathbf{z}_n)$ on the r.h.s. of (13.68) Moreover $p(\ldots)$ should be $\ln p(\ldots)$ on the r.h.s. of (13.70).

We start by rewriting (13.6) as

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) \prod_{n=2}^{N} p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1}).$$

Taking the logarithm we get

$$\ln p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = \ln p(\mathbf{z}_1) + \ln p(\mathbf{x}_1|\mathbf{z}_1) + \sum_{n=2}^{N} (\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}))$$

where, with the first two terms we have recovered the r.h.s. of (13.69). We now use this to maximize over $\mathbf{z}_1, \ldots, \mathbf{z}_N$,

$$\begin{aligned}
&\max_{\mathbf{z}_1,\ldots,\mathbf{z}_N} \left\{ \omega(\mathbf{z}_1) + \sum_{n=2}^{N} [\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n|\mathbf{z}_{n-1})] \right\} \\
&= \max_{\mathbf{z}_2,\ldots,\mathbf{z}_N} \left\{ \ln p(\mathbf{x}_2|\mathbf{z}_2) + \max_{\mathbf{z}_1} \{\ln p(\mathbf{z}_2|\mathbf{z}_1) + \omega(\mathbf{z}_1)\} \right. \\
&\qquad \left. + \sum_{n=3}^{N} [\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n|\mathbf{z}_{n-1})] \right\} \\
&= \max_{\mathbf{z}_2,\ldots,\mathbf{z}_N} \left\{ \omega(\mathbf{z}_2) + \sum_{n=3}^{N} [\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n|\mathbf{z}_{n-1})] \right\} \qquad (325)
\end{aligned}$$

where we have exchanged the order of maximization and summation for $\mathbf{z}_2$ to recover (13.68) for $n = 2$, and since the first and the last line of (325) have identical forms, this extends recursively to all $n > 2$.

13.17 The emission probabilities over observed variables $\mathbf{x}_n$ are absorbed into the corresponding factors, f_n, analogously to the way in which Figure 13.14 was transformed into Figure 13.15. The factors then take the form

$$h(\mathbf{z}_1) = p(\mathbf{z}_1|\mathbf{u}_1)p(\mathbf{x}_1|\mathbf{z}_1, \mathbf{u}_1) \qquad (326)$$

$$f_n(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{u}_n)p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{u}_n). \qquad (327)$$

13.18 By combining the results from Solution 13.17 with those from Section 13.2.3, the desired outcome is easily obtained.

By combining (327) with (13.49) and (13.50), we see that

$$\alpha(\mathbf{z}_n) = \sum_{\mathbf{z}_{n-1}} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{u}_n)p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{u}_n)\alpha(\mathbf{z}_{n-1})$$

corresponding to (13.36). The initial condition is given directly by (326) and corresponds to (13.37).

Similarly, from (327), (13.51) and (13.52), we see that

$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} p(\mathbf{z}_{n+1}|\mathbf{z}_n, \mathbf{u}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}, \mathbf{u}_{n+1})\beta(\mathbf{z}_{n+1})$$

which corresponds to (13.38). The presence of the input variables does not affect the initial condition $\beta(\mathbf{z}_N) = 1$.

13.19 Since the joint distribution over all variables, latent and observed, is Gaussian, we can maximize w.r.t. any chosen set of variables. In particular, we can maximize w.r.t. all the latent variables jointly or maximize each of the marginal distributions separately. However, from (2.98), we see that the resulting means will be the same in both cases and since the mean and the mode coincide for the Gaussian, maximizing w.r.t. to latent variables jointly and individually will yield the same result.

13.20 Making the following substitions from the l.h.s. of (13.87),

$$\mathbf{x} \Rightarrow \mathbf{z}_{n-1} \quad \boldsymbol{\mu} \Rightarrow \boldsymbol{\mu}_{n-1} \quad \boldsymbol{\Lambda}^{-1} \Rightarrow \mathbf{V}_{n-1}$$
$$\mathbf{y} \Rightarrow \mathbf{z}_n \quad \mathbf{A} \Rightarrow \mathbf{A} \quad \mathbf{b} \Rightarrow \mathbf{0} \quad \mathbf{L}^{-1} \Rightarrow \boldsymbol{\Gamma},$$

in (2.113) and (2.114), (2.115) becomes

$$p(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\mathbf{A}\boldsymbol{\mu}_{n-1}, \boldsymbol{\Gamma} + \mathbf{A}\mathbf{V}_{n-1}\mathbf{A}^{\mathrm{T}}),$$

as desired.

13.21 If we substitute the r.h.s. of (13.87) for the integral on the r.h.s. of (13.86), we get

$$c_n\mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}_n|\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{P}_{n-1}).$$

The r.h.s. define the joint probability distribution over $\mathbf{x}_n$ and $\mathbf{z}_n$ in terms of a conditional distribution over $\mathbf{x}_n$ given $\mathbf{z}_n$ and a distribution over $\mathbf{z}_n$, corresponding to (2.114) and (2.113), respectively. What we need to do is to rewrite this into a conditional distribution over $\mathbf{z}_n$ given $\mathbf{x}_n$ and a distribution over $\mathbf{x}_n$, corresponding to (2.116) and (2.115), respectively.

If we make the substitutions

$$\mathbf{x} \Rightarrow \mathbf{z}_n \quad \boldsymbol{\mu} \Rightarrow \mathbf{A}\boldsymbol{\mu}_{n-1} \quad \boldsymbol{\Lambda}^{-1} \Rightarrow \mathbf{P}_{n-1}$$
$$\mathbf{y} \Rightarrow \mathbf{x}_n \quad \mathbf{A} \Rightarrow \mathbf{C} \quad \mathbf{b} \Rightarrow \mathbf{0} \quad \mathbf{L}^{-1} \Rightarrow \boldsymbol{\Sigma},$$

in (2.113) and (2.114), (2.115) directly gives us the r.h.s. of (13.91).

From (2.114), we have that

$$p(\mathbf{z}_n|\mathbf{x}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n) = \mathcal{N}(\mathbf{z}_n|\mathbf{M}(\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + \mathbf{P}_{n-1}^{-1}\mathbf{A}\boldsymbol{\mu}_{n-1}), \mathbf{M}), \quad (328)$$

where we have used (2.117) to define

$$\mathbf{M} = (\mathbf{P}_{n-1} + \mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}. \quad (329)$$

Using (C.7) and (13.92), we can rewrite (329) as follows:

$$\begin{aligned}
\mathbf{M} &= (\mathbf{P}_{n-1} + \mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1} \\
&= \mathbf{P}_{n-1} - \mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}}(\boldsymbol{\Sigma} + \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}})^{-1}\mathbf{C}\mathbf{P}_{n-1} \\
&= (\mathbf{I} - \mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}}(\boldsymbol{\Sigma} + \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}})^{-1}\mathbf{C})\mathbf{P}_{n-1} \\
&= (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{P}_{n-1},
\end{aligned}$$

which equals the r.h.s. of (13.90).

Using (329), (C.5) and (13.92), we can derive the following equality:

$$\begin{aligned}\mathbf{M}\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} &= (\mathbf{P}_{n-1}+\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} \\ &= \mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}}(\mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}}+\boldsymbol{\Sigma})^{-1} = \mathbf{K}_n.\end{aligned}$$

Using this and (13.90), we can rewrite the expression for the mean in (328) as follows:

$$\begin{aligned}\mathbf{M}(\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}_n+\mathbf{P}_{n-1}^{-1}\mathbf{A}\boldsymbol{\mu}_{n-1}) &= \mathbf{M}\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}_n+(\mathbf{I}-\mathbf{K}_n\mathbf{C})\mathbf{A}\boldsymbol{\mu}_{n-1} \\ &= \mathbf{K}_n\mathbf{x}_n+\mathbf{A}\boldsymbol{\mu}_{n-1}-\mathbf{K}_n\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1} \\ &= \mathbf{A}\boldsymbol{\mu}_{n-1}+\mathbf{K}_n(\mathbf{x}_n-\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}),\end{aligned}$$

which equals the r.h.s. of (13.89).

13.22 Using (13.76), (13.77) and (13.84), we can write (13.93), for the case $n=1$, as

$$c_1\mathcal{N}(\mathbf{z}_1|\boldsymbol{\mu}_1,\mathbf{V}_1)=\mathcal{N}(\mathbf{z}_1|\boldsymbol{\mu}_0,\mathbf{V}_0)\mathcal{N}(\mathbf{x}_1|\mathbf{C}\mathbf{z}_1,\boldsymbol{\Sigma}).$$

The r.h.s. define the joint probability distribution over $\mathbf{x}_1$ and $\mathbf{z}_1$ in terms of a conditional distribution over $\mathbf{x}_1$ given $\mathbf{z}_1$ and a distribution over $\mathbf{z}_1$, corresponding to (2.114) and (2.113), respectively. What we need to do is to rewrite this into a conditional distribution over $\mathbf{z}_1$ given $\mathbf{x}_1$ and a distribution over $\mathbf{x}_1$, corresponding to (2.116) and (2.115), respectively.

If we make the substitutions

$$\mathbf{x}\Rightarrow\mathbf{z}_1 \quad \boldsymbol{\mu}\Rightarrow\boldsymbol{\mu}_0 \quad \boldsymbol{\Lambda}^{-1}\Rightarrow\mathbf{V}_0$$

$$\mathbf{y}\Rightarrow\mathbf{x}_1 \quad \mathbf{A}\Rightarrow\mathbf{C} \quad \mathbf{b}\Rightarrow\mathbf{0} \quad \mathbf{L}^{-1}\Rightarrow\boldsymbol{\Sigma},$$

in (2.113) and (2.114), (2.115) directly gives us the r.h.s. of (13.96).

13.23 Using (13.76) and (13.77) we can rewrite (13.93) as

$$c_1\widehat{\alpha}(\mathbf{z}_1)=\mathcal{N}\left(\mathbf{z}_1|\boldsymbol{\mu}_0,\mathbf{V}_0\right)\mathcal{N}\left(\mathbf{x}_1|\mathbf{C}\mathbf{z}_1,\boldsymbol{\Sigma}\right).$$

Making the same substitutions as in Solution 13.22, (2.115) and (13.96) give

$$p(\mathbf{x}_1)=\mathcal{N}\left(\mathbf{x}_1|\mathbf{C}\boldsymbol{\mu}_0,\boldsymbol{\Sigma}+\mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\right)=c_1.$$

Hence, from the product rule and (2.116),

$$\widehat{\alpha}(\mathbf{z}_1)=p(\mathbf{z}_1|\mathbf{x}_1)=\mathcal{N}\left(\mathbf{z}_1|\boldsymbol{\mu}_1,\mathbf{V}_1\right)$$

where, from (13.97) and (C.7),

$$\begin{aligned}\mathbf{V}_1 &= \left(\mathbf{V}_0^{-1}+\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{C}\right)^{-1} \\ &= \mathbf{V}_0-\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\left(\boldsymbol{\Sigma}+\mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\right)^{-1}\mathbf{C}\mathbf{V}_0 \\ &= \left(\mathbf{I}-\mathbf{K}_1\mathbf{C}\right)\mathbf{V}_0\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\mu}_1 &= \mathbf{V}_1\left(\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}_1 + \mathbf{V}_0^{-1}\boldsymbol{\mu}_0\right)\\ &= \boldsymbol{\mu}_0 + \mathbf{K}_1\left(\mathbf{x}_1 - \mathbf{C}\boldsymbol{\mu}_0\right)\end{aligned}$$

where we have used

$$\begin{aligned}\mathbf{V}_1\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} &= \mathbf{V}_0\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} - \mathbf{K}_1\mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\\ &= \mathbf{V}_0\mathbf{C}^{\mathrm{T}}\left(\mathbf{I} - \left(\boldsymbol{\Sigma} + \mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\right)^{-1}\mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\right)\boldsymbol{\Sigma}^{-1}\\ &= \mathbf{V}_0\mathbf{C}^{\mathrm{T}}\Big(\mathbf{I} - \left(\boldsymbol{\Sigma} + \mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\right)^{-1}\mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\\ &\qquad + \left(\boldsymbol{\Sigma} + \mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\right)^{-1}\boldsymbol{\Sigma} - \left(\boldsymbol{\Sigma} + \mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\right)^{-1}\boldsymbol{\Sigma}\Big)\boldsymbol{\Sigma}^{-1}\\ &= \mathbf{V}_0\mathbf{C}^{\mathrm{T}}\left(\boldsymbol{\Sigma} + \mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}}\right)^{-1} = \mathbf{K}_1.\end{aligned}$$

13.24 This extension can be embedded in the existing framework by adopting the following modifications:

$$\boldsymbol{\mu}_0' = \begin{bmatrix}\boldsymbol{\mu}_0\\ 1\end{bmatrix} \quad \mathbf{V}_0' = \begin{bmatrix}\mathbf{V}_0 & \mathbf{0}\\ \mathbf{0} & 0\end{bmatrix} \quad \boldsymbol{\Gamma}' = \begin{bmatrix}\boldsymbol{\Gamma} & \mathbf{0}\\ \mathbf{0} & 0\end{bmatrix}$$

$$\mathbf{A}' = \begin{bmatrix}\mathbf{A} & \mathbf{a}\\ \mathbf{0} & 1\end{bmatrix} \quad \mathbf{C}' = \begin{bmatrix}\mathbf{C} & \mathbf{c}\end{bmatrix}.$$

This will ensure that the constant terms $\mathbf{a}$ and $\mathbf{c}$ are included in the corresponding Gaussian means for $\mathbf{z}_n$ and $\mathbf{x}_n$ for $n = 1, \ldots, N$.

Note that the resulting covariances for $\mathbf{z}_n$, $\mathbf{V}_n$, will be singular, as will the corresponding prior covariances, $\mathbf{P}_{n-1}$. This will, however, only be a problem where these matrices need to be inverted, such as in (13.102). These cases must be handled separately, using the 'inversion' formula

$$(\mathbf{P}'_{n-1})^{-1} = \begin{bmatrix}\mathbf{P}_{n-1}^{-1} & \mathbf{0}\\ \mathbf{0} & 0\end{bmatrix},$$

nullifying the contribution from the (non-existent) variance of the element in $\mathbf{z}_n$ that accounts for the constant terms $\mathbf{a}$ and $\mathbf{c}$.

13.25 **NOTE**: In PRML, the second half on the third sentence in the exercise should read: "...in which $\mathbf{C} = 1$, $\mathbf{A} = 1$ and $\boldsymbol{\Gamma} = 0$." Moreover, in the following sentence $\mathbf{m}_0$ and $\mathbf{V}_0$ should be replaced by $\boldsymbol{\mu}_0$ and $\mathbf{P}_0$; see also the PRML errata.

Since $\mathbf{C} = 1$, $\mathbf{P}_0 = \sigma_0^2$ and $\boldsymbol{\Sigma} = \sigma^2$, (13.97) gives

$$\mathbf{K}_1 = \frac{\sigma_0^2}{\sigma_0^2 + \sigma}.$$

Substituting this into (13.94) and (13.95), we get

$$\begin{aligned}
\mu_1 &= \mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\sigma}(x_1-\mu_0) \\
&= \frac{1}{\sigma_0^2+\sigma}(\sigma_0^2 x_1 + \sigma\mu_0) \\
\sigma_1^2 &= \left(1 - \frac{\sigma_0^2}{\sigma_0^2+\sigma}\right)\sigma_0^2 \\
&= \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma}
\end{aligned}$$

where σ_1^2 replaces $\mathbf{V}_1$. We note that these agree with (2.141) and (2.142), respectively.

We now assume that (2.141) and (2.142) hold for N, and we rewrite them as

$$\mu_N = \sigma_N^2\left(\frac{1}{\sigma_0^2}\mu_0 + \frac{N}{\sigma^2}\mu_{\mathrm{ML}}^{(N)}\right) \tag{330}$$

$$\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma} \tag{331}$$

where, analogous to (2.143),

$$\mu_{\mathrm{ML}}^{(N)} = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{332}$$

Since $\mathbf{A} = 1$ and $\mathbf{\Gamma} = 0$, (13.88) gives

$$\mathbf{P}_N = \sigma_N^2 \tag{333}$$

substituting this into (13.92), we get

$$\mathbf{K}_{N+1} = \frac{\sigma_N^2}{N\sigma_N^2+\sigma} \tag{334}$$

Using (331), (333), (334) and (13.90), we get

$$\begin{aligned}
\sigma_{N+1}^2 &= \left(1 - \frac{\sigma_N^2}{\sigma_N^2+\sigma}\right)\sigma_N^2 \\
&= \frac{\sigma_N^2\sigma^2}{\sigma_N^2+\sigma} \\
&= \frac{\sigma_0^2\sigma^4/(\sigma_0^2+\sigma)}{(\sigma^2\sigma_0^2\sigma^4 + \sigma^2 N\sigma_0^2)/(\sigma_0^2+\sigma)} \\
&= \frac{\sigma_0^2\sigma^2}{(N+1)\sigma_0^2+\sigma}.
\end{aligned} \tag{335}$$

Using (330), (332), (334), (335) and (13.89), we get

$$\begin{aligned}
\mu_{N+1} &= \mu_N + \frac{\sigma_N^2}{\sigma_N^2+\sigma}(x_{N+1}-\mu_N) \\
&= \frac{1}{\sigma_N^2+\sigma}(\sigma_N^2 x_{N+1} + \sigma\mu_N) \\
&= \frac{\sigma_N^2}{\sigma_0^2+\sigma}x_{N+1} + \frac{\sigma_N^2\sigma^2}{\sigma_N^2+\sigma}\left(\frac{1}{\sigma_0^2}\mu_0 + \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n\right) \\
&= \sigma_{N+1}^2\left(\frac{1}{\sigma_0^2}\mu_0 + \frac{N+1}{\sigma^2}\mu_{\mathrm{ML}}^{(N+1)}\right).
\end{aligned}$$

Thus (330) and (331) must hold for all $N \geqslant 1$.

13.26 **NOTE**: In the 1st printing of PRML, equation (12.42) contains a mistake; the covariance on the r.h.s. should be $\sigma^2\mathbf{M}^{-1}$. Furthermore, the exercise should make explicit the assumption that $\boldsymbol{\mu} = \mathbf{0}$ in (12.42).

From (13.84) and the assumption that $\mathbf{A} = \mathbf{0}$, we have

$$p(\mathbf{z}_n|\mathbf{x}_1,\ldots,\mathbf{x}_n) = p(\mathbf{z}_n|\mathbf{x}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n) \tag{336}$$

where $\boldsymbol{\mu}_n$ and $\mathbf{V}_n$ are given by (13.89) and (13.90), respectively. Since $\mathbf{A} = \mathbf{0}$ and $\boldsymbol{\Gamma} = \mathbf{I}$, $\mathbf{P}_{n-1} = \mathbf{I}$ for all n and thus (13.92) becomes

$$\begin{aligned}
\mathbf{K}_n &= \mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}}\left(\mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1} \\
&= \mathbf{W}^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}\right)^{-1}
\end{aligned} \tag{337}$$

where we have substituted $\mathbf{W}$ for $\mathbf{C}$ and $\sigma^2\mathbf{I}$ for $\boldsymbol{\Sigma}$. Using (337), (12.41), (C.6) and (C.7), (13.89) and (13.90) can be rewritten as

$$\begin{aligned}
\boldsymbol{\mu}_n &= \mathbf{K}_n\mathbf{x}_n \\
&= \mathbf{W}^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_n \\
&= \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{x}_n
\end{aligned}$$

$$\begin{aligned}
\mathbf{V}_n &= (\mathbf{I} - \mathbf{K}_n\mathbf{C})\,\mathbf{P}_{n-1} = \mathbf{I} - \mathbf{K}_n\mathbf{W} \\
&= \mathbf{I} - \mathbf{W}^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{W} \\
&= \left(\sigma^{-2}\mathbf{W}^{\mathrm{T}}\mathbf{W} + \mathbf{I}\right)^{-1} \\
&= \sigma^2\left(\mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I}\right)^{-1} = \sigma^2\mathbf{M}^{-1}
\end{aligned}$$

which makes (336) equivalent with (12.42), assuming $\boldsymbol{\mu} = \mathbf{0}$.

13.27 **NOTE**: In the 1st printing of PRML, this exercise should have made explicit the assumption that $\mathbf{C} = \mathbf{I}$ in (13.86).

From (13.86), it is easily seen that if $\mathbf{\Sigma}$ goes to $\mathbf{0}$, the posterior over $\mathbf{z}_n$ will become completely determined by $\mathbf{x}_n$, since the first factor on the r.h.s. of (13.86), and hence also the l.h.s., will collapse to a spike at $\mathbf{x}_n = \mathbf{C}\mathbf{z}_n$.

13.28 **NOTE**: In PRML, this exercise should also assume that $\mathbf{C} = \mathbf{I}$., Moreover, $\mathbf{V}_0$ should be replaced by $\mathbf{P}_0$ in the text of the exercise; see also the PRML errata.

Starting from (13.75) and (13.77), we can use (2.113)–(2.117) to obtain

$$p(\mathbf{z}_1|\mathbf{x}_1) = \mathcal{N}\left(\mathbf{z}_1|\boldsymbol{\mu}_1, \mathbf{V}_1\right)$$

where

$$\boldsymbol{\mu}_1 = \mathbf{V}_1\left(\mathbf{C}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x}_1 + \mathbf{P}_0^{-1}\boldsymbol{\mu}_0\right) = \mathbf{x}_1 \tag{338}$$

$$\mathbf{V}_1 = \left(\mathbf{P}_0^{-1} + \mathbf{C}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{C}\right)^{-1} = \mathbf{\Sigma} \tag{339}$$

since $\mathbf{P}_0 \to \infty$ and $\mathbf{C} = \mathbf{I}$; note that these results can equally well be obtained from (13.94), (13.95) and (13.97).

Now we assume that for N

$$\boldsymbol{\mu}_N = \overline{\mathbf{x}}_N = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{340}$$

$$\mathbf{V}_N = \frac{1}{N}\mathbf{\Sigma} \tag{341}$$

and we note that these assumptions are met by (338) and (339), respectively. From (13.88) and (341), we then have

$$\mathbf{P}_N = \mathbf{V}_N = \frac{1}{N}\mathbf{\Sigma} \tag{342}$$

since $\mathbf{C} = \mathbf{I}$ and $\mathbf{\Gamma} = \mathbf{0}$. Using this together with (13.92), we obtain

$$\begin{aligned}
\mathbf{K}_{N+1} &= \mathbf{P}_N\mathbf{C}^{\mathrm{T}}\left(\mathbf{C}\mathbf{P}_N\mathbf{C}^{\mathrm{T}} + \mathbf{\Sigma}\right)^{-1} \\
&= \mathbf{P}_N\left(\mathbf{P}_N + \mathbf{\Sigma}\right)^{-1} \\
&= \frac{1}{N}\mathbf{\Sigma}\left(\frac{1}{N}\mathbf{\Sigma} + \mathbf{\Sigma}\right)^{-1} \\
&= \frac{1}{N}\mathbf{\Sigma}\left(\frac{N+1}{N}\mathbf{\Sigma}\right)^{-1} \\
&= \frac{1}{N+1}\mathbf{I}.
\end{aligned}$$

Substituting this into (13.89) and (13.90), making use of (340) and (342), we have

$$\begin{aligned}
\boldsymbol{\mu}_{N+1} &= \boldsymbol{\mu}_N + \frac{1}{N+1}(\mathbf{x}_{N+1} - \boldsymbol{\mu}_N) \\
&= \overline{\mathbf{x}}_N + \frac{1}{N+1}(\mathbf{x}_{N+1} - \overline{\mathbf{x}}_N) \\
&= \frac{1}{N+1}\mathbf{x}_{N+1} + \left(1 - \frac{1}{N+1}\right)\frac{1}{N}\sum_{n=1}^{N} \\
&= \frac{1}{N+1}\sum_{n=1}^{N+1}\mathbf{x}_n = \overline{\mathbf{x}}_{N+1} \\
\mathbf{V}_{N+1} &= \left(\mathbf{I} - \frac{1}{N+1}\mathbf{I}\right)\frac{1}{N}\boldsymbol{\Sigma} \\
&= \frac{1}{N+1}\boldsymbol{\Sigma}.
\end{aligned}$$

Thus, (340) and (341) holds for all $N \geqslant 1$.

13.29 **NOTE**: In the 1st printing of PRML, $\boldsymbol{\mu}_N$ should be $\boldsymbol{\mu}_n$ on the r.h.s. of (13.100)

Multiplying both sides of (13.99) by $\widehat{\alpha}(\mathbf{z}_n)$, and then making use of (13.98), we get

$$c_{n+1}\mathcal{N}\left(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n\right) = \widehat{\alpha}(\mathbf{z}_n)\int \widehat{\beta}(\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n)\,\mathrm{d}\mathbf{z}_{n+1} \tag{343}$$

Using (2.113)–(2.117), (13.75) and (13.84), we have

$$\begin{aligned}
\widehat{\alpha}(\mathbf{z}_n)p(\mathbf{z}_{n+1}|\mathbf{z}_n) &= \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n)\,\mathcal{N}(\mathbf{z}_{n+1}|\mathbf{A}\mathbf{z}_n, \boldsymbol{\Gamma}) \\
&= \mathcal{N}(\mathbf{z}_{n+1}|\mathbf{A}\boldsymbol{\mu}_n, \mathbf{A}\mathbf{V}_n\mathbf{A} + \boldsymbol{\Gamma})\,\mathcal{N}(\mathbf{z}_n|\mathbf{m}_n, \mathbf{M}_n) \qquad (344)
\end{aligned}$$

where

$$\mathbf{m}_n = \mathbf{M}_n\left(\mathbf{A}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{z}_{n+1} + \mathbf{V}_n^{-1}\boldsymbol{\mu}_n\right) \tag{345}$$

and, using (C.7) and (13.102),

$$\begin{aligned}
\mathbf{M}_n &= \left(\mathbf{A}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{A} + \mathbf{V}_n\right)^{-1} \qquad (346) \\
&= \mathbf{V}_n - \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\left(\boldsymbol{\Gamma} + \mathbf{A}\mathbf{V}_n\mathbf{A}^{\mathrm{T}}\right)^{-1}\mathbf{A}\mathbf{V}_n \\
&= \mathbf{V}_n - \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\mathbf{P}_n^{-1}\mathbf{A}\mathbf{V}_n \qquad (347) \\
&= \left(\mathbf{I} - \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\mathbf{P}_n^{-1}\mathbf{A}\right)\mathbf{V}_n \\
&= (\mathbf{I} - \mathbf{J}_n\mathbf{A})\,\mathbf{V}_n \qquad (348)
\end{aligned}$$

Substituting the r.h.s. of (344) into (343) and then making use of (13.85)–(13.88) and

(13.98), we have

$$\begin{aligned} c_{n+1}\mathcal{N}\left(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n\right) &= \int \widehat{\beta}(\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})\mathcal{N}\left(\mathbf{z}_{n+1}|\mathbf{A}\boldsymbol{\mu}_n, \mathbf{P}_n\right) \\ &\qquad \mathcal{N}\left(\mathbf{z}_n|\mathbf{m}_n, \mathbf{M}_n\right)\,\mathrm{d}\mathbf{z}_{n+1} \\ &= \int \widehat{\beta}(\mathbf{z}_{n+1})c_{n+1}\widehat{\alpha}(\mathbf{z}_{n+1})\mathcal{N}\left(\mathbf{z}_n|\mathbf{m}_n, \mathbf{M}_n\right)\,\mathrm{d}\mathbf{z}_{n+1} \\ &= c_{n+1}\int \gamma(\mathbf{z}_{n+1})\mathcal{N}\left(\mathbf{z}_n|\mathbf{m}_n, \mathbf{M}_n\right)\,\mathrm{d}\mathbf{z}_{n+1} \\ &= c_{n+1}\int \mathcal{N}\left(\mathbf{z}_{n+1}|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n\right)\mathcal{N}\left(\mathbf{z}_n|\mathbf{m}_n, \mathbf{M}_n\right)\,\mathrm{d}\mathbf{z}_{n+1}. \end{aligned}$$

Thus, from this, (345) and (2.113)–(2.115), we see that

$$\widehat{\boldsymbol{\mu}}_n = \mathbf{M}_n\left(\mathbf{A}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\widehat{\boldsymbol{\mu}}_{n+1} + \mathbf{V}_n^{-1}\boldsymbol{\mu}_n\right) \tag{349}$$

$$\widehat{\mathbf{V}}_n = \mathbf{M}_n\mathbf{A}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\widehat{\mathbf{V}}_{n+1}\boldsymbol{\Gamma}^{-1}\mathbf{A}\mathbf{M}_n + \mathbf{M}_n. \tag{350}$$

Using (347) and (13.102), we see that

$$\begin{aligned} \mathbf{M}_n\mathbf{A}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1} &= \left(\mathbf{V} - \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\mathbf{P}_n^{-1}\mathbf{A}\mathbf{V}_n\right)\mathbf{A}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1} \\ &= \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}_n^{-1}\mathbf{A}\mathbf{V}_n\mathbf{A}^{\mathrm{T}}\right)\boldsymbol{\Gamma}^{-1} \\ &= \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}_n^{-1}\mathbf{A}\mathbf{V}_n\mathbf{A}^{\mathrm{T}} - \mathbf{P}_n^{-1}\boldsymbol{\Gamma} + \mathbf{P}_n^{-1}\boldsymbol{\Gamma}\right)\boldsymbol{\Gamma}^{-1} \\ &= \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}_n^{-1}\mathbf{P}_n + \mathbf{P}_n^{-1}\boldsymbol{\Gamma}\right)\boldsymbol{\Gamma}^{-1} \\ &= \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\mathbf{P}_n^{-1} = \mathbf{J}_n \end{aligned} \tag{351}$$

and using (348) together with (351), we can rewrite (349) as (13.100). Similarly, using (13.102), (347) and (351), we rewrite (350) as

$$\begin{aligned} \widehat{\mathbf{V}}_n &= \mathbf{M}_n\mathbf{A}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\widehat{\mathbf{V}}_{n+1}\boldsymbol{\Gamma}^{-1}\mathbf{A}\mathbf{M}_n + \mathbf{M}_n \\ &= \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathrm{T}} + \mathbf{V}_n - \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\mathbf{P}_n^{-1}\mathbf{A}\mathbf{V}_n \\ &= \mathbf{V}_n + \mathbf{J}_n\left(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n\right)\mathbf{J}_n^{\mathrm{T}} \end{aligned}$$

13.30 **NOTE**: See note in Solution 13.15.

The first line of (13.103) corresponds exactly to (13.65). We then use (13.75), (13.76), (13.84) and (13.98) to rewrite $p(\mathbf{z}_n|\mathbf{z}_{n-1})$, $p(\mathbf{x}_n|\mathbf{z}_n)$, $\widehat{\alpha}(\mathbf{z}_{n-1})$ and $\widehat{\beta}(\mathbf{z}_n)$, respectively, yielding the second line of (13.103).

13.31 Substituting the r.h.s. of (13.84) for $\widehat{\alpha}(\mathbf{z}_n)$ in (13.103) and then using (2.113)–(2.117) and (13.86), we get

$$\begin{aligned}
&\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) \\
&= \frac{\mathcal{N}\left(\mathbf{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \mathbf{V}_{n-1}\right)\mathcal{N}\left(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \boldsymbol{\Gamma}\right)\mathcal{N}\left(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma}\right)\mathcal{N}\left(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n\right)}{\mathcal{N}\left(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n\right)c_n} \\
&= \frac{\mathcal{N}\left(\mathbf{z}_n|\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{P}_{n-1}\right)\mathcal{N}\left(\mathbf{z}_{n-1}|\mathbf{m}_{n-1}, \mathbf{M}_{n-1}\right)\mathcal{N}\left(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n\right)}{\mathcal{N}\left(\mathbf{z}_n|\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{P}_{n-1}\right)} \\
&= \mathcal{N}\left(\mathbf{z}_{n-1}|\mathbf{m}_{n-1}, \mathbf{M}_{n-1}\right)\mathcal{N}\left(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n\right) \qquad (352)
\end{aligned}$$

where $\mathbf{m}_{n-1}$ and $\mathbf{M}_{n-1}$ are given by (345) and (346). Equation (13.104) then follows from (345), (351) and (352).

13.32 **NOTE**: In PRML, $\mathbf{V}_0$ should be replaced by $\mathbf{P}_0$ in the text of the exercise; see also the PRML errata.

We can write the expected complete log-likelihood, given by the equation after (13.109), as a function of $\boldsymbol{\mu}_0$ and $\mathbf{P}_0$, as follows:

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) &= -\frac{1}{2}\ln|\mathbf{P}_0| \\
&\quad -\frac{1}{2}\mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\text{old}}}\left[\mathbf{z}_1^{\mathrm{T}}\mathbf{P}_0^{-1}\mathbf{z}_1 - \mathbf{z}_1^{\mathrm{T}}\mathbf{P}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^{\mathrm{T}}\mathbf{P}_0^{-1}\mathbf{z}_1 + \boldsymbol{\mu}_0^{\mathrm{T}}\mathbf{P}_0^{-1}\boldsymbol{\mu}_0\right] \qquad (353) \\
&= \frac{1}{2}\left(\ln|\mathbf{P}_0^{-1}| - \mathrm{Tr}\left[\mathbf{P}_0^{-1}\mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\text{old}}}\left[\mathbf{z}_1\mathbf{z}_1^{\mathrm{T}} - \mathbf{z}_1\boldsymbol{\mu}_0^{\mathrm{T}} - \boldsymbol{\mu}_0\mathbf{z}_1^{\mathrm{T}} + \boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathrm{T}}\right]\right]\right), \qquad (354)
\end{aligned}$$

where we have used (C.13) and omitted terms independent of $\boldsymbol{\mu}_0$ and $\mathbf{P}_0$.

From (353), we can calculate the derivative w.r.t. $\boldsymbol{\mu}_0$ using (C.19), to get

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_0} = 2\mathbf{P}_0^{-1}\boldsymbol{\mu}_0 - 2\mathbf{P}_0^{-1}\mathbb{E}[\mathbf{z}_1].$$

Setting this to zero and rearranging, we immediately obtain (13.110).

Using (354), (C.24) and (C.28), we can evaluate the derivatives w.r.t. $\mathbf{P}_0^{-1}$,

$$\frac{\partial Q}{\partial \mathbf{P}_0^{-1}} = \frac{1}{2}\left(\mathbf{P}_0 - \mathbb{E}[\mathbf{z}_1\mathbf{z}_1^{\mathrm{T}}] - \mathbb{E}[\mathbf{z}_1]\boldsymbol{\mu}_0^{\mathrm{T}} - \boldsymbol{\mu}_0\mathbb{E}[\mathbf{z}_1^{\mathrm{T}}] + \boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathrm{T}}\right).$$

Setting this to zero, rearrangning and making use of (13.110), we get (13.111).

13.33 **NOTE**: In PRML, the first instance of $\mathbf{A}^{\text{new}}$ on the second line of equation (13.114) should be transposed.

Expanding the square in the second term of (13.112) and making use of the trace operator, we obtain

$$\mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}}\left[\frac{1}{2}\sum_{n=2}^{N}\left(\mathbf{z}_n-\mathbf{A}\mathbf{z}_{n-1}\right)^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\left(\mathbf{z}_n-\mathbf{A}\mathbf{z}_{n-1}\right)\right] = \mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}}\left[\frac{1}{2}\sum_{n=2}^{N}\mathrm{Tr}\left(\boldsymbol{\Gamma}^{-1}\left\{\mathbf{A}\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} - \mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\mathbf{z}_{n-1}\mathbf{z}_n^{\mathrm{T}} + \mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right\}\right)\right]. \quad (355)$$

Using results from Appendix C, we can calculate the derivative of this w.r.t. $\mathbf{A}$, yielding

$$\frac{\partial Q}{\partial \mathbf{A}} = \boldsymbol{\Gamma}^{-1}\mathbf{A}\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right] - \boldsymbol{\Gamma}^{-1}\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right].$$

Setting this equal to zero and solving for $\mathbf{A}$, we obtain (13.113).

Using (355) and results from Appendix C, we can calculate the derivative of (13.112) w.r.t. $\boldsymbol{\Gamma}^{-1}$, to obtain

$$\frac{\partial Q}{\partial \boldsymbol{\Gamma}} = \frac{N-1}{2}\boldsymbol{\Gamma} - \frac{1}{2}\sum_{n=2}^{N}\left(\mathbf{A}\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right]\mathbf{A}^{\mathrm{T}} - \mathbb{E}\left[\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right]\mathbf{A}^{\mathrm{T}} - \mathbf{A}\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_n^{\mathrm{T}}\right] + \mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right]\right).$$

Setting this equal to zero, substituting $\mathbf{A}^{\mathrm{new}}$ for $\mathbf{A}$ and solving for $\boldsymbol{\Gamma}$, we obtain (13.114).

13.34 **NOTE**: In PRML, the first and third instances of $\mathbf{C}^{\mathrm{new}}$ on the second line of equation (13.116) should be transposed.

By making use of (C.28), equations (13.115) and (13.116) are obtained in an identical manner to (13.113) and (13.114), respectively, in Solution 13.33.

Chapter 14 Combining Models

14.1 The required predictive distribution is given by

$$p(\mathbf{t}|\mathbf{x},\mathbf{X},\mathbf{T}) = \sum_h p(h)\sum_{\mathbf{z}_h} p(\mathbf{z}_h)\int p(\mathbf{t}|\mathbf{x},\boldsymbol{\theta}_h,\mathbf{z}_h,h)p(\boldsymbol{\theta}_h|\mathbf{X},\mathbf{T},h)\,\mathrm{d}\boldsymbol{\theta}_h, \quad (356)$$

where

$$\begin{aligned} p(\boldsymbol{\theta}_h|\mathbf{X},\mathbf{T},h) &= \frac{p(\mathbf{T}|\mathbf{X},\boldsymbol{\theta}_h,h)p(\boldsymbol{\theta}_h|h)}{p(\mathbf{T}|\mathbf{X},h)} \\ &\propto p(\boldsymbol{\theta}|h)\prod_{n=1}^{N} p(\mathbf{t}_n|\mathbf{x}_n,\boldsymbol{\theta},h) \\ &= p(\boldsymbol{\theta}|h)\prod_{n=1}^{N}\left(\sum_{\mathbf{z}_{nh}} p(\mathbf{t}_n,\mathbf{z}_{nh}|\mathbf{x}_n,\boldsymbol{\theta},h)\right) \end{aligned} \tag{357}$$

The integrals and summations in (356) are examples of Bayesian averaging, accounting for the uncertainty about which model, h, is the correct one, the value of the corresponding parameters, $\boldsymbol{\theta}_h$, and the state of the latent variable, $\mathbf{z}_h$. The summation in (357), on the other hand, is an example of the use of latent variables, where different data points correspond to different latent variable states, although all the data are assumed to have been generated by a single model, h.

14.2 Using (14.13), we can rewrite (14.11) as

$$\begin{aligned} E_{\mathrm{COM}} &= \mathbb{E}_{\mathbf{x}}\left[\left\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\right\}^2\right] \\ &= \frac{1}{M^2}\mathbb{E}_{\mathbf{x}}\left[\left\{\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\right\}^2\right] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\mathbb{E}_{\mathbf{x}}\left[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})\right] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}\left[\epsilon_m(\mathbf{x})^2\right] = \frac{1}{M}E_{\mathrm{AV}} \end{aligned}$$

where we have used (14.10) in the last step.

14.3 We start by rearranging the r.h.s. of (14.10), by moving the factor $1/M$ inside the sum and the expectation operator outside the sum, yielding

$$\mathbb{E}_{\mathbf{x}}\left[\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2\right].$$

If we then identify $\epsilon_m(\mathbf{x})$ and $1/M$ with x_i and λ_i in (1.115), respectively, and take $f(x) = x^2$, we see from (1.115) that

$$\left(\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})\right)^2 \leqslant \sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2.$$

Since this holds for all values of $\mathbf{x}$, it must also hold for the expectation over $\mathbf{x}$, proving (14.54).

14.4 If $E(y(\mathbf{x}))$ is convex, we can apply (1.115) as follows:

$$\begin{aligned} E_{\text{AV}} &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}\left[E(y(\mathbf{x}))\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\sum_{m=1}^{M}\frac{1}{M}E(y(\mathbf{x}))\right] \\ &\geqslant \mathbb{E}_{\mathbf{x}}\left[E\left(\sum_{m=1}^{M}\frac{1}{M}y(\mathbf{x})\right)\right] \\ &= E_{\text{COM}} \end{aligned}$$

where $\lambda_i = 1/M$ for $i = 1, \ldots, M$ in (1.115) and we have implicitly defined versions of E_{AV} and E_{COM} corresponding to $E(y(\mathbf{x}))$.

14.5 To prove that (14.57) is a sufficient condition for (14.56) we have to show that (14.56) follows from (14.57). To do this, consider a fixed set of $y_m(\mathbf{x})$ and imagine varying the α_m over all possible values allowed by (14.57) and consider the values taken by $y_{\text{COM}}(\mathbf{x})$ as a result. The maximum value of $y_{\text{COM}}(\mathbf{x})$ occurs when $\alpha_k = 1$ where $y_k(\mathbf{x}) \geqslant y_m(\mathbf{x})$ for $m \neq k$, and hence all $\alpha_m = 0$ for $m \neq k$. An analogous result holds for the minimum value. For other settings of $\boldsymbol{\alpha}$,

$$y_{\min}(\mathbf{x}) < y_{\text{COM}}(\mathbf{x}) < y_{\max}(\mathbf{x}),$$

since $y_{\text{COM}}(\mathbf{x})$ is a convex combination of points, $y_m(\mathbf{x})$, such that

$$\forall m : y_{\min}(\mathbf{x}) \leqslant y_m(\mathbf{x}) \leqslant y_{\max}(\mathbf{x}).$$

Thus, (14.57) is a sufficient condition for (14.56).

Showing that (14.57) is a necessary condition for (14.56) is equivalent to showing that (14.56) is a sufficient condition for (14.57). The implication here is that if (14.56) holds for any choice of values of the committee members $\{y_m(\mathbf{x})\}$ then (14.57) will be satisfied. Suppose, without loss of generality, that α_k is the smallest of the α values, i.e. $\alpha_k \leqslant \alpha_m$ for $k \neq m$. Then consider $y_k(\mathbf{x}) = 1$, together with $y_m(\mathbf{x}) = 0$ for all $m \neq k$. Then $y_{\min}(\mathbf{x}) = 0$ while $y_{\text{COM}}(\mathbf{x}) = \alpha_k$ and hence from (14.56) we obtain $\alpha_k \geqslant 0$. Since α_k is the smallest of the α values it follows that all of the coefficients must satisfy $\alpha_k \geqslant 0$. Similarly, consider the case in which $y_m(\mathbf{x}) = 1$ for all m. Then $y_{\min}(\mathbf{x}) = y_{\max}(\mathbf{x}) = 1$, while $y_{\text{COM}}(\mathbf{x}) = \sum_m \alpha_m$. From (14.56) it then follows that $\sum_m \alpha_m = 1$, as required.

14.6 If we differentiate (14.23) w.r.t. α_m we obtain

$$\frac{\partial E}{\partial \alpha_m} = \frac{1}{2}\left((e^{\alpha_m/2} + e^{-\alpha_m/2})\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n) - e^{-\alpha_m/2}\sum_{n=1}^{N} w_n^{(m)}\right).$$

Setting this equal to zero and rearranging, we get

$$\frac{\sum_n w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_n w_n^{(m)}} = \frac{e^{-\alpha_m/2}}{e^{\alpha_m/2} + e^{-\alpha_m/2}} = \frac{1}{e^{\alpha_m} + 1}.$$

Using (14.16), we can rewrite this as

$$\frac{1}{e^{\alpha_m} + 1} = \epsilon_m,$$

which can be further rewritten as

$$e^{\alpha_m} = \frac{1 - \epsilon_m}{\epsilon_m},$$

from which (14.17) follows directly.

14.7 Taking the functional derivative of (14.27) w.r.t. $y(\mathbf{x})$, we get

$$\begin{aligned}\frac{\delta}{\delta y(\mathbf{x})}\mathbb{E}_{\mathbf{x},t}\left[\exp\left\{-ty(\mathbf{x})\right\}\right] &= -\sum_t t \exp\left\{-ty(\mathbf{x})\right\} p(t|\mathbf{x})p(\mathbf{x}) \\ &= \left\{\exp\left\{y(\mathbf{x})\right\} p(t=-1|\mathbf{x}) - \exp\left\{-y(\mathbf{x})\right\} p(t=+1|\mathbf{x})\right\} p(\mathbf{x}).\end{aligned}$$

Setting this equal to zero and rearranging, we obtain (14.28).

14.8 Assume that (14.20) is a negative log likelihood function. Then the corresponding likelihood function is given by

$$\exp(-E) = \prod_{n=1}^{N} \exp\left(-\exp\left\{-t_n f_m(\mathbf{x}_n)\right\}\right)$$

and thus

$$p(t_n|\mathbf{x}_n) \propto \exp\left(-\exp\left\{-t_n f_m(\mathbf{x}_n)\right\}\right).$$

We can normalize this probability distribution by computing the normalization constant

$$Z = \exp\left(-\exp\left\{f_m(\mathbf{x}_n)\right\}\right) + \exp\left(-\exp\left\{-f_m(\mathbf{x}_n)\right\}\right)$$

but since Z involves $f_m(\mathbf{x})$, the log of the resulting normalized probability distribution no longer corresponds to (14.20) as a function of $f_m(\mathbf{x})$.

14.9 The sum-of-squares error for the additive model of (14.21) is defined as

$$E = \frac{1}{2}\sum_{n=1}^{N}(t_n - f_m(\mathbf{x}_n))^2.$$

Using (14.21), we can rewrite this as

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - f_{m-1}(\mathbf{x}_n) - \frac{1}{2}\alpha_m y_m(\mathbf{x}))^2,$$

where we recognize the two first terms inside the square as the residual from the $(m-1)$-th model. Minimizing this error w.r.t. $y_m(\mathbf{x})$ will be equivalent to fitting $y_m(\mathbf{x})$ to the (scaled) residuals.

14.10 The error function that we need to minimize is

$$E(t) = \frac{1}{2}\sum_{\{t_n\}}(t_n - t)^2 .$$

Taking the derivative of this w.r.t. t and setting it equal to zero we get

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\sum_{\{t_n\}}(t_n - t) = 0.$$

Solving for t yields

$$t = \frac{1}{N'}\sum_{\{t_n\}} t_n$$

where $N' = |\{t_n\}|$, i.e. the number of values in $\{t_n\}$.

14.11 **NOTE**: In PRML, the text of this exercise contains mistakes; please refer to the PRML Errata for relevant corrections.

The misclassification rates for the two tree models are given by

$$\begin{aligned} R_{\mathrm{A}} &= \frac{100+100}{400+400} = \frac{1}{4} \\ R_{\mathrm{B}} &= \frac{0+200}{400+400} = \frac{1}{4} \end{aligned}$$

From (14.31) and (14.32) we see that the pruning criterion for the cross-entropy case evaluates to

$$\begin{aligned} C_{\mathrm{Xent}}(T_{\mathrm{A}}) &= -2\left(\frac{100}{400}\ln\frac{100}{400} + \frac{300}{400}\ln\frac{300}{400}\right) + 2\lambda \simeq 1.12 + 2\lambda \\ C_{\mathrm{Xent}}(T_{\mathrm{B}}) &= -\frac{400}{400}\ln\frac{400}{400} - \frac{200}{400}\ln\frac{200}{400} - \frac{0}{400}\ln\frac{0}{400} - \frac{200}{400}\ln\frac{200}{400} + 2\lambda \\ &\simeq 0.69 + 2\lambda \end{aligned}$$

Finally, from (14.31) and (14.33) we see that the pruning criterion for the Gini index case become

$$\begin{aligned} C_{\mathrm{Gini}}(T_{\mathrm{A}}) &= 2\left[\frac{300}{400}\left(1 - \frac{300}{400}\right) + \frac{100}{400}\left(1 - \frac{100}{400}\right)\right] + 2\lambda = \frac{3}{4} + 2\lambda \\ C_{\mathrm{Gini}}(T_{\mathrm{B}}) &= \frac{400}{400}\left(1 - \frac{400}{400}\right) + \frac{200}{400}\left(1 - \frac{200}{400}\right) \\ &\quad + \frac{0}{400}\left(1 - \frac{0}{400}\right) + \frac{200}{400}\left(1 - \frac{200}{400}\right) + 2\lambda = \frac{1}{2} + 2\lambda. \end{aligned}$$

Thus we see that, while both trees have the same misclassification rate, B performs better in terms of cross-entropy as well as Gini index.

14.12 Drawing on (3.32), we redefine (14.34) as

$$p(\mathbf{t}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{t}|\mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}, \beta^{-1}\mathbf{I}\right)$$

and then make the corresponding changes to (14.35)–(14.37) and $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}\right)$; also $\mathbf{t}$ will be replaced by $\mathbf{T}$ to align with the notation used in Section 3.1.5. Equation (14.39) will now take the form

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}\right) = \sum_{n=1}^{N} \gamma_{nk} \left\{ -\frac{\beta}{2} \left\| \mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}_n \right\|^2 \right\} + \text{const.}$$

Following the same steps as in the single target case, we arrive at a corresponding version of (14.42):

$$\mathbf{W}_k = \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}_k\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}_k\mathbf{T}.$$

For β, (14.43) becomes

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}\right) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left\{ \frac{D}{2} \ln \beta - \frac{\beta}{2} \left\| \mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}_n \right\|^2 \right\}$$

and consequently (14.44) becomes

$$\frac{1}{\beta} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left\| \mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}_n \right\|^2.$$

14.13 Starting from the mixture distribution in (14.34), we follow the same steps as for mixtures of Gaussians, presented in Section 9.2. We introduce a K-nomial latent variable, $\mathbf{z}$, such that the joint distribution over $\mathbf{z}$ and t equals

$$p(t, \mathbf{z}) = p(t|\mathbf{z})p(\mathbf{z}) = \prod_{k=1}^{K} \left(\mathcal{N}\left(t|\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}, \beta^{-1}\right) \pi_k \right)^{z_k}.$$

Given a set of observations, $\{(t_n, \boldsymbol{\phi}_n)\}_{n=1}^{N}$, we can write the complete likelihood over these observations and the corresponding $\mathbf{z}_1, \ldots, \mathbf{z}_N$, as

$$\prod_{n=1}^{N} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(t_n|\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}) \right)^{z_{nk}}.$$

Taking the logarithm, we obtain (14.36).

14.14 Since $Q(\theta, \theta^{\text{old}})$ (defined by the unnumbered equation preceeding (14.38)) has exactly the same dependency on $\boldsymbol{\pi}$ as (9.40), (14.38) can be derived just like the corresponding result in Solution 9.9.

14.15 The predictive distribution from the mixture of linear regression models for a new input feature vector, $\widehat{\boldsymbol{\phi}}$, is obtained from (14.34), with $\boldsymbol{\phi}$ replaced by $\widehat{\boldsymbol{\phi}}$. Calculating the expectation of t under this distribution, we obtain

$$\mathbb{E}[t|\widehat{\boldsymbol{\phi}}, \boldsymbol{\theta}] = \sum_{k=1}^{K} \pi_k \mathbb{E}[t|\widehat{\boldsymbol{\phi}}, \mathbf{w}_k, \beta].$$

Depending on the parameters, this expectation is potentially K-modal, with one mode for each mixture component. However, the weighted combination of these modes output by the mixture model may not be close to any single mode. For example, the combination of the two modes in the left panel of Figure 14.9 will end up in between the two modes, a region with no signicant probability mass.

14.16 This solution is analogous to Solution 14.12. It makes use of results from Section 4.3.4 in the same way that Solution 14.12 made use of results from Section 3.1.5. Note, however, that Section 4.3.4 uses k as class index and K to denote the number of classes, whereas here we will use c and C, respectively, for these purposes. This leaves k and K for mixture component indexing and number of mixture components, as used elsewhere in Chapter 14.

Using 1-of-C coding for the targets, we can look to (4.107) to rewrite (14.45) as

$$p(\mathbf{t}|\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{c=1}^{C} y_{kc}^{t_c}$$

and making the corresponding changes to (14.46)–(14.48), which lead to an expected complete-data log likelihood function,

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left\{ \ln \pi_k + \sum_{c=1}^{C} t_{nc} \ln y_{nkc} \right\}$$

corresponding to (14.49).

As in the case of the mixture of logistic regression models, the M step for $\boldsymbol{\pi}$ is the same as for other mixture models, given by (14.50). In the M step for $\mathbf{W}_1, \ldots, \mathbf{W}_K$, where

$$\mathbf{W}_k = [\mathbf{w}_{k1}, \ldots, \mathbf{w}_{kC}]$$

we can again deal with each mixture component separately, using an iterative method such as IRLS, to solve

$$\nabla_{\mathbf{w}_{kc}} Q = \sum_{n=1}^{N} \gamma_{nk} \left(y_{nkc} - t_{nc}\right) \boldsymbol{\phi}_n = 0$$

where we have used (4.109) and (14.51). We obtain the corresponding Hessian from (4.110) (**NOTE**: In the 1st printing of PRML, the leading minus sign on the r.h.s. should be removed.) and (14.52) as

$$\mathbf{H}_k = \nabla_{\mathbf{w}_{kc}} \nabla_{\mathbf{w}_{k\hat{c}}} Q = \sum_{n=1}^{N} \gamma_{nk} y_{nkc} \left(I_{c\hat{c}} - y_{nk\hat{c}} \right) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}.$$

14.17 If we define $\psi_k(t|\mathbf{x})$ in (14.58) as

$$\psi_k(t|\mathbf{x}) = \sum_{m=1}^{M} \lambda_{mk} \phi_{mk}(t|\mathbf{x}),$$

we can rewrite (14.58) as

$$\begin{aligned} p(t|\mathbf{x}) &= \sum_{k=1}^{K} \pi_k \sum_{m=1}^{M} \lambda_{mk} \phi_{mk}(t|\mathbf{x}) \\ &= \sum_{k=1}^{K} \sum_{m=1}^{M} \pi_k \lambda_{mk} \phi_{mk}(t|\mathbf{x}). \end{aligned}$$

By changing the indexation, we can write this as

$$p(t|\mathbf{x}) = \sum_{l=1}^{L} \eta_l \phi_l(t|\mathbf{x}),$$

where $L = KM$, $l = (k-1)M + m$, $\eta_l = \pi_k \lambda_{mk}$ and $\phi_l(\cdot) = \phi_{mk}(\cdot)$. By construction, $\eta_l \geqslant 0$ and $\sum_{l=1}^{L} \eta_l = 1$.

Note that this would work just as well if π_k and λ_{mk} were to be dependent on $\mathbf{x}$, as long as they both respect the constraints of being non-negative and summing to 1 for every possible value of $\mathbf{x}$.

Finally, consider a tree-structured, hierarchical mixture model, as illustrated in the left panel of Figure 12. On the top (root) level, this is a mixture with two components. The mixing coefficients are given by a linear logistic regression model and hence are input dependent. The left sub-tree correspond to a local conditional density model, $\psi_1(t|\mathbf{x})$. In the right sub-tree, the structure from the root is replicated, with the difference that both sub-trees contain local conditional density models, $\psi_2(t|\mathbf{x})$ and $\psi_3(t|\mathbf{x})$.

We can write the resulting mixture model on the form (14.58) with mixing coefficients

$$\begin{aligned} \pi_1(\mathbf{x}) &= \sigma(\mathbf{v}_1^{\mathrm{T}}\mathbf{x}) \\ \pi_2(\mathbf{x}) &= (1 - \sigma(\mathbf{v}_1^{\mathrm{T}}\mathbf{x}))\sigma(\mathbf{v}_2^{\mathrm{T}}\mathbf{x}) \\ \pi_3(\mathbf{x}) &= (1 - \sigma(\mathbf{v}_1^{\mathrm{T}}\mathbf{x}))(1 - \sigma(\mathbf{v}_2^{\mathrm{T}}\mathbf{x})), \end{aligned}$$

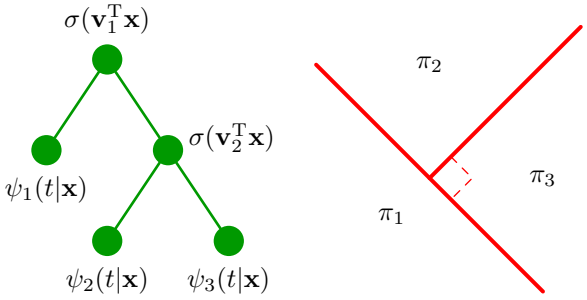


Figure 12 Left: an illustration of a hierarchical mixture model, where the input dependent mixing coefficients are determined by linear logistic models associated with interior nodes; the leaf nodes correspond to local (conditional) density models. Right: a possible division of the input space into regions where different mixing coefficients dominate, under the model illustrated left.

where $\sigma(\cdot)$ is defined in (4.59) and $\mathbf{v}_1$ and $\mathbf{v}_2$ are the parameter vectors of the logistic regression models. Note that $\pi_1(\mathbf{x})$ is independent of the value of $\mathbf{v}_2$. This would not be the case if the mixing coefficients were modelled using a single level softmax model,

$$\pi_k(\mathbf{x}) = \frac{e^{\mathbf{u}_k^{\mathrm{T}}\mathbf{x}}}{\sum_j^3 e^{\mathbf{u}_j^{\mathrm{T}}\mathbf{x}}},$$

where the parameters $\mathbf{u}_k$, corresponding to $\pi_k(\mathbf{x})$, will also affect the other mixing coeffiecients, $\pi_{j\neq k}(\mathbf{x})$, through the denominator. This gives the hierarchical model different properties in the modelling of the mixture coefficients over the input space, as compared to a linear softmax model. An example is shown in the right panel of Figure 12, where the red lines represent borders of equal mixing coefficients in the input space. These borders are formed from two straight lines, corresponding to the two logistic units in the left panel of 12. A corresponding division of the input space by a softmax model would involve three straight lines joined at a single point, looking, e.g., something like the red lines in Figure 4.3 in PRML; note that a linear three-class softmax model could not implement the borders show in right panel of Figure 12.