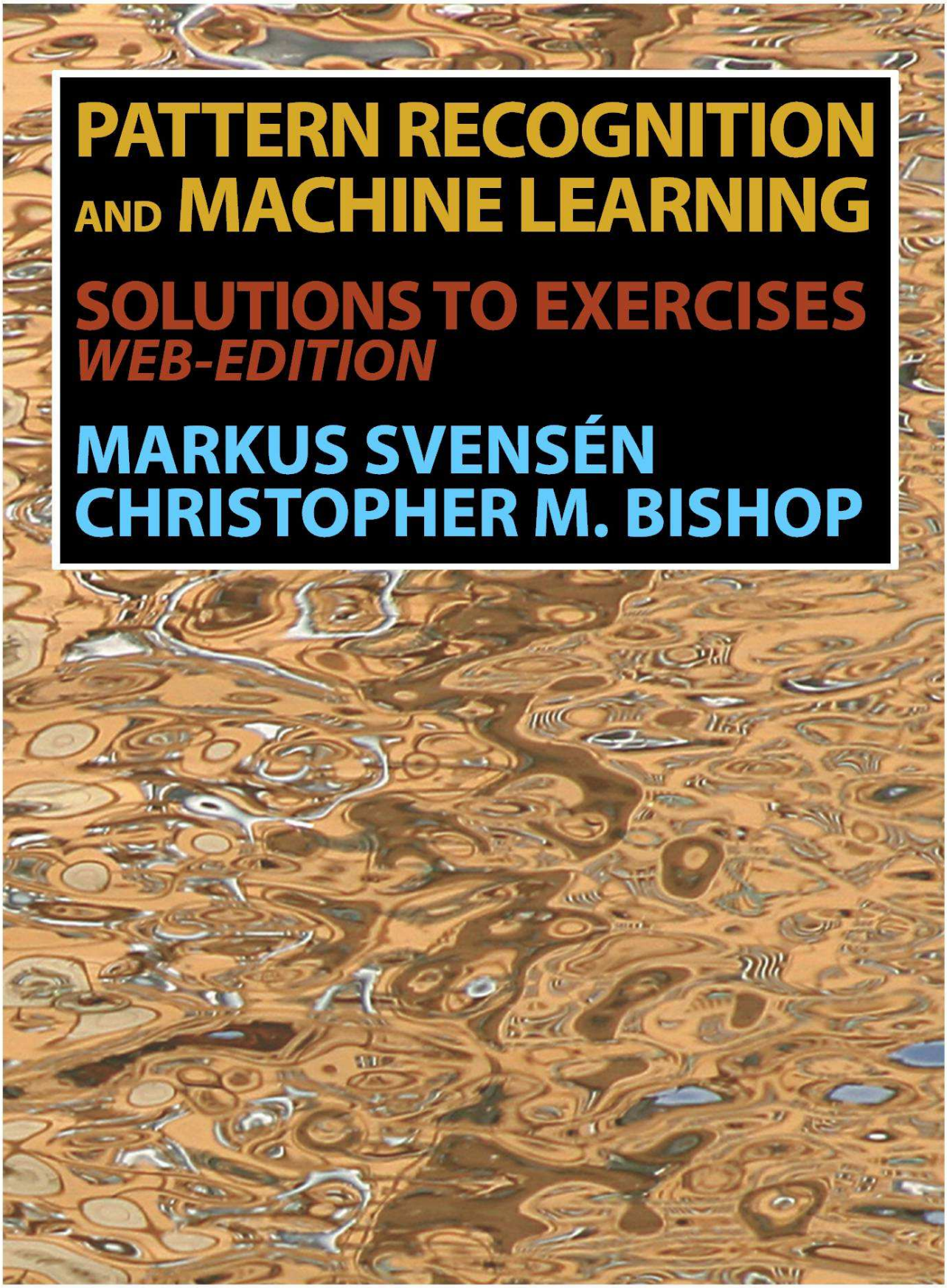

PATTERN RECOGNITION
AND MACHINE LEARNING
SOLUTIONS TO EXERCISES
WEB-EDITION
MARKUS SVENSÉN
CHRISTOPHER M. BISHOP

Pattern Recognition and Machine Learning
Solutions to the Exercises: Web-Edition

Markus Svensén and Christopher M. Bishop

This is the solutions manual (web-edition) for the book *Pattern Recognition and Machine Learning* (PRML; published by Springer in 2006). It contains solutions to the WWW exercises. This release was created August 3, 2007; eventual future releases with corrections to errors will be published on the PRML web-site (see below).

The authors would like to express their gratitude to the various people who have provided feedback on pre-releases of this document. In particular, the "Bishop Reading Group", held in the Visual Geometry Group at the University of Oxford provided valuable comments and suggestions.

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 Markus Svensén, markussv@microsoft.com.

Further information about PRML is available from:

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

Contents

Chapter 1 Pattern Recognition

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.

1.4 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

$$\widetilde{f}(y) = f(g(y)). \tag{2}$$

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 (2) with respect to y

$$\widetilde{f}'(\widehat{y}) = f'(g(\widehat{y}))g'(\widehat{y}) = 0. \tag{3}$$

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

$$p'_y(y) = sp'_x(g(y))\{g'(y)\}^2 + sp_x(g(y))g''(y). \tag{4}$$

Due to the presence of the second term on the right hand side of (4) 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

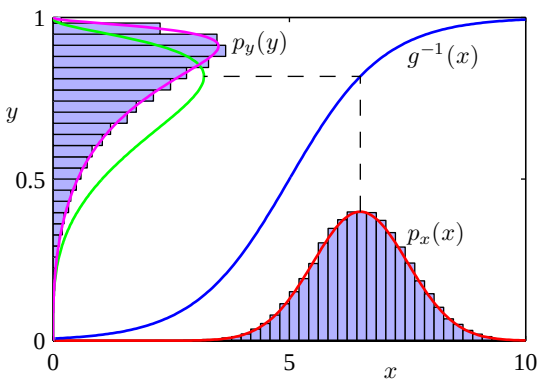


Figure 1 Example of the transformation of the mode of a density under a non-linear change of variables, illustrating the different behaviour compared to a simple function. See the text for details.

the right hand side of (4) 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

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

The inverse of this function is given by

$$y = g^{-1}(x) = \frac{1}{1+\exp(-x+5)} \tag{6}$$

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 (6), 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!

1.7 The transformation from Cartesian to polar coordinates is defined by

$$x = r\cos\theta \tag{7}$$
$$y = r\sin\theta \tag{8}$$

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

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

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

$$I^2 = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) r \,\mathrm{d}r\,\mathrm{d}\theta \tag{9}$$

$$= 2\pi \int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2}\right)\frac{1}{2}\,\mathrm{d}u \tag{10}$$

$$= \pi\left[\exp\left(-\frac{u}{2\sigma^2}\right)(-2\sigma^2)\right]_0^{\infty} \tag{11}$$

$$= 2\pi\sigma^2 \tag{12}$$

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

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

$$= \frac{I}{(2\pi\sigma^2)^{1/2}} = 1$$

as required.

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

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}x\,\mathrm{d}x. \tag{13}$$

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

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}y^2\right\}(y+\mu)\,\mathrm{d}y. \tag{14}$$

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

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \mathrm{d}x = \left(2\pi\sigma^2\right)^{1/2}. \tag{15}$$

We now differentiate both sides of (15) with respect to σ^2 and then re-arrange to obtain

$$\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} (x-\mu)^2 \,\mathrm{d}x = \sigma^2 \tag{16}$$

which directly shows that

$$\mathbb{E}[(x-\mu)^2] = \mathrm{var}[x] = \sigma^2. \tag{17}$$

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) 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 $\mathbf{x} = \boldsymbol{\mu}$.

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

$$\begin{aligned}\mathbb{E}[x+z] &= \iint (x+z)p(x)p(z)\,\mathrm{d}x\,\mathrm{d}z &(18)\\ &= \int xp(x)\,\mathrm{d}x + \int zp(z)\,\mathrm{d}z &(19)\\ &= \mathbb{E}[x] + \mathbb{E}[z]. &(20)\end{aligned}$$

Similarly for the variances, we first note that

$$(x+z-\mathbb{E}[x+z])^2 = (x-\mathbb{E}[x])^2 + (z-\mathbb{E}[z])^2 + 2(x-\mathbb{E}[x])(z-\mathbb{E}[z]) \quad (21)$$

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

$$\begin{aligned} \mathrm{var}[x+z] &= \iint (x+z-\mathbb{E}[x+z])^2 p(x)p(z)\,\mathrm{d}x\,\mathrm{d}z \\ &= \int (x-\mathbb{E}[x])^2 p(x)\,\mathrm{d}x + \int (z-\mathbb{E}[z])^2 p(z)\,\mathrm{d}z \\ &= \mathrm{var}(x) + \mathrm{var}(z). \end{aligned} \quad (22)$$

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

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

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[x_n] = \mu \quad (23)$$

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

$$\begin{aligned} \mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(x_n - \frac{1}{N}\sum_{m=1}^{N} x_m\right)^2\right] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - \frac{2}{N}x_n\sum_{m=1}^{N}x_m + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}x_m x_l\right] \\ &= \left\{\mu^2 + \sigma^2 - 2\left(\mu^2 + \frac{1}{N}\sigma^2\right) + \mu^2 + \frac{1}{N}\sigma^2\right\} \\ &= \left(\frac{N-1}{N}\right)\sigma^2 \end{aligned} \quad (24)$$

as required.

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

$$n(D, M) = \sum_{i_1=1}^{D} \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 \tag{25}$$

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

$$n(D, M) = \sum_{i_1=1}^{D} \left\{ \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 \right\} \tag{26}$$

where the term in braces has $M-1$ terms which, from (25), must equal $n(i_1, M-1)$. Thus we can write

$$n(D, M) = \sum_{i_1=1}^{D} n(i_1, M-1) \tag{27}$$

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

$$\begin{aligned} \sum_{i=1}^{D+1} \frac{(i+M-2)!}{(i-1)!(M-1)!} &= \frac{(D+M-1)!}{(D-1)!M!} + \frac{(D+M-1)!}{D!(M-1)!} \\ &= \frac{(D+M-1)!D + (D+M-1)!M}{D!M!} \\ &= \frac{(D+M)!}{D!M!} \end{aligned} \tag{28}$$

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

$$n(D, M-1) = \frac{(D+M-2)!}{(D-1)!\,(M-1)!}. \tag{29}$$

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

$$n(D, M) = \sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!\,(M-1)!} \tag{30}$$

which, making use of (1.136), gives

$$n(D,M) = \frac{(D+M-1)!}{(D-1)!\,M!} \tag{31}$$

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.

1.17 Using integration by parts we have

$$\begin{aligned}\Gamma(x+1) &= \int_0^\infty u^x e^{-u}\,\mathrm{d}u \\ &= \left[-e^{-u}u^x\right]_0^\infty + \int_0^\infty xu^{x-1}e^{-u}\,\mathrm{d}u = 0 + x\Gamma(x).\end{aligned} \tag{32}$$

For $x = 1$ we have

$$\Gamma(1) = \int_0^\infty e^{-u}\,\mathrm{d}u = \left[-e^{-u}\right]_0^\infty = 1. \tag{33}$$

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 (32) 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

$$\frac{1}{2}S_D \int_0^\infty e^{-u}u^{D/2-1}\,\mathrm{d}u = \frac{1}{2}S_D\Gamma(D/2) \tag{34}$$

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

$$V_D = S_D \int_0^1 r^{D-1}\,\mathrm{d}r = \frac{S_D}{D}. \tag{35}$$

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

$$S_2 = 2\pi, \qquad S_3 = 4\pi, \qquad V_2 = \pi a^2, \qquad V_3 = \frac{4}{3}\pi a^3. \tag{36}$$

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

$$\int_{\text{shell}} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \simeq p(r)S_D r^{D-1}\epsilon \tag{37}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}r}p(r) \propto \left[(D-1)r^{D-2} + r^{D-1}\left(-\frac{r}{\sigma^2}\right)\right]\exp\left(-\frac{r^2}{2\sigma^2}\right) = 0. \tag{38}$$

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

Next we note that

$$\begin{aligned} p(\widehat{r}+\epsilon) &\propto (\widehat{r}+\epsilon)^{D-1}\exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2} + (D-1)\ln(\widehat{r}+\epsilon)\right]. \end{aligned} \tag{39}$$

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 $\exp(D/2)$.

1.22 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.

1.24 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

$$j = \arg\min_l \sum_k L_{kl}p(\mathcal{C}_k|\mathbf{x}) \tag{40}$$

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

$$\text{choose} \begin{cases} \text{class } j, & \text{if } \min_l \sum_k L_{kl}p(\mathcal{C}_k|\mathbf{x}) < \lambda; \\ \text{reject}, & \text{otherwise.} \end{cases} \tag{41}$$

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).

1.27 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

$$\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x}) \, \mathrm{d}t \tag{42}$$

for each value of $\mathbf{x}$. Setting the derivative of (42) 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

$$\int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1} p(t|\mathbf{x}) \, \mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1} p(t|\mathbf{x}) \, \mathrm{d}t. \tag{43}$$

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

$$\int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x})\,\mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x})\,\mathrm{d}t. \tag{44}$$

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 (42) 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 (42) by choosing the location of the notch to coincide with the largest value of $p(t)$, i.e. with the (conditional) mode.

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

$$H(x) = -\sum_{i=1}^{M} p(x_i) \ln p(x_i) = \sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)}. \tag{45}$$

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

$$H(x) \leqslant \ln\left(\sum_{i=1}^{M} p(x_i)\frac{1}{p(x_i)}\right) = \ln M. \tag{46}$$

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.

1.34 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

$$\lambda_1 = 1 - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \tag{47}$$

$$\lambda_2 = 0 \tag{48}$$

$$\lambda_3 = \frac{1}{2\sigma^2} \tag{49}$$

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 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).

1.38 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)

$$f\left(\sum_{i=1}^{M+1}\lambda_i x_i\right) = f\left(\lambda_{M+1}x_{M+1} + \sum_{i=1}^{M}\lambda_i x_i\right) \tag{50}$$

$$= f\left(\lambda_{M+1}x_{M+1} + (1-\lambda_{M+1})\sum_{i=1}^{M}\eta_i x_i\right) \tag{51}$$

where we have defined

$$\eta_i = \frac{\lambda_i}{1-\lambda_{M+1}}. \tag{52}$$

We now apply (1.114) to give

$$f\left(\sum_{i=1}^{M+1}\lambda_i x_i\right) \leqslant \lambda_{M+1}f(x_{M+1}) + (1-\lambda_{M+1})f\left(\sum_{i=1}^{M}\eta_i x_i\right). \tag{53}$$

We now note that the quantities λ_i by definition satisfy

$$\sum_{i=1}^{M+1}\lambda_i = 1 \tag{54}$$

and hence we have

$$\sum_{i=1}^{M} \lambda_i = 1 - \lambda_{M+1} \tag{55}$$

Then using (52) we see that the quantities η_i satisfy the property

$$\sum_{i=1}^{M} \eta_i = \frac{1}{1-\lambda_{M+1}} \sum_{i=1}^{M} \lambda_i = 1. \tag{56}$$

Thus we can apply the result (1.115) at order M and so (53) becomes

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leqslant \lambda_{M+1} f(x_{M+1}) + (1-\lambda_{M+1}) \sum_{i=1}^{M} \eta_i f(x_i) = \sum_{i=1}^{M+1} \lambda_i f(x_i) \tag{57}$$

where we have made use of (52).

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

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &= -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= -\int p(\mathbf{y}) \ln p(\mathbf{y}) \, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \end{aligned} \tag{58}$$

Chapter 2 Density Estimation

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

$$\sum_{x \in \{0,1\}} p(x|\mu) = p(x=0|\mu) + p(x=1|\mu) \tag{59}$$

$$= (1-\mu) + \mu = 1 \tag{60}$$

$$\sum_{x \in \{0,1\}} x p(x|\mu) = 0.p(x=0|\mu) + 1.p(x=1|\mu) = \mu \tag{61}$$

$$\sum_{x \in \{0,1\}} (x-\mu)^2 p(x|\mu) = \mu^2 p(x=0|\mu) + (1-\mu)^2 p(x=1|\mu) \tag{62}$$

$$= \mu^2(1-\mu) + (1-\mu)^2\mu = \mu(1-\mu). \tag{63}$$

The entropy is given by

$$\begin{aligned} H(x) &= -\sum_{x\in\{0,1\}} p(x|\mu)\ln p(x|\mu) \\ &= -\sum_{x\in\{0,1\}} \mu^x(1-\mu)^{1-x}\{x\ln\mu + (1-x)\ln(1-\mu)\} \\ &= -(1-\mu)\ln(1-\mu) - \mu\ln\mu. \end{aligned} \tag{64}$$

2.3 Using the definition (2.10) we have

$$\begin{aligned} \binom{N}{n} + \binom{N}{n-1} &= \frac{N!}{n!(N-n)!} + \frac{N!}{(n-1)!(N+1-n)!} \\ &= \frac{(N+1-n)N! + nN!}{n!(N+1-n)!} = \frac{(N+1)!}{n!(N+1-n)!} \\ &= \binom{N+1}{n}. \end{aligned} \tag{65}$$

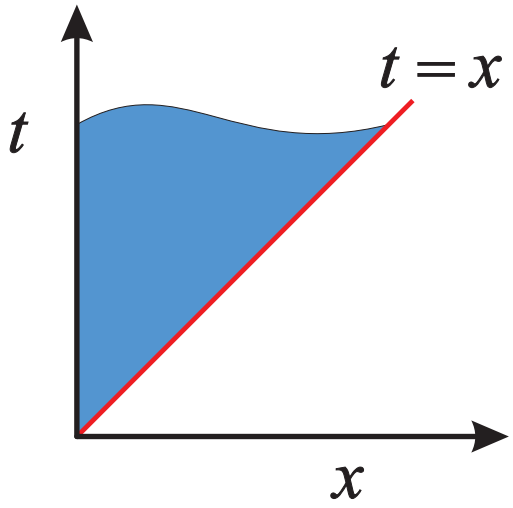
To prove the binomial theorem (2.263) we note that the theorem is trivially true for $N = 0$. We now assume that it holds for some general value N and prove its correctness for $N + 1$, which can be done as follows

$$\begin{aligned} (1+x)^{N+1} &= (1+x)\sum_{n=0}^{N}\binom{N}{n}x^n \\ &= \sum_{n=0}^{N}\binom{N}{n}x^n + \sum_{n=1}^{N+1}\binom{N}{n-1}x^n \\ &= \binom{N}{0}x^0 + \sum_{n=1}^{N}\left\{\binom{N}{n} + \binom{N}{n-1}\right\}x^n + \binom{N}{N}x^{N+1} \\ &= \binom{N+1}{0}x^0 + \sum_{n=1}^{N}\binom{N+1}{n}x^n + \binom{N+1}{N+1}x^{N+1} \\ &= \sum_{n=0}^{N+1}\binom{N+1}{n}x^n \end{aligned} \tag{66}$$

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

$$\begin{aligned} \sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n} &= (1-\mu)^N\sum_{n=0}^{N}\binom{N}{n}\left(\frac{\mu}{1-\mu}\right)^n \\ &= (1-\mu)^N\left(1+\frac{\mu}{1-\mu}\right)^N = 1 \end{aligned} \tag{67}$$

Figure 2 Plot of the region of integration of (68) in (x, t) space.



as required.

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

$$\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1} \left\{ \int_x^\infty \exp(-t)(t-x)^{b-1} \, \mathrm{d}t \right\} \mathrm{d}x. \tag{68}$$

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

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^t x^{a-1} \exp(-t)(t-x)^{b-1} \, \mathrm{d}x \, \mathrm{d}t. \tag{69}$$

The change in the limits of integration in going from (68) to (69) can be understood by reference to Figure 2. Finally we change variables in the x integral using $x = t\mu$ to give

$$\begin{aligned} \Gamma(a)\Gamma(b) &= \int_0^\infty \exp(-t) t^{a-1} t^{b-1} t \, \mathrm{d}t \int_0^1 \mu^{a-1}(1-\mu)^{b-1} \, \mathrm{d}\mu \\ &= \Gamma(a+b) \int_0^1 \mu^{a-1}(1-\mu)^{b-1} \, \mathrm{d}\mu. \end{aligned} \tag{70}$$

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

$$p_{M-1}(\mu_1, \ldots, \mu_{M-2}) = \int_0^{1-\sum_{j=1}^{M-2}\mu_j} p_M(\mu_1, \ldots, \mu_{M-1}) \, \mathrm{d}\mu_{M-1}$$

$$= C_M \left[\prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \right] \int_0^{1-\sum_{j=1}^{M-2}\mu_j} \mu_{M-1}^{\alpha_{M-1}-1} \left(1 - \sum_{j=1}^{M-1} \mu_j \right)^{\alpha_M - 1} \mathrm{d}\mu_{M-1}.$$

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

$$\mu_{M-1} = t\left(1 - \sum_{j=1}^{M-2} \mu_j\right) \tag{71}$$

which gives

$$\begin{aligned} & p_{M-1}(\mu_1, \ldots, \mu_{M-2}) \\ &= C_M \left[\prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1}\right] \left(1 - \sum_{j=1}^{M-2} \mu_j\right)^{\alpha_{M-1}+\alpha_M - 1} \int_0^1 t^{\alpha_{M-1}-1}(1-t)^{\alpha_M - 1}\,\mathrm{d}t \\ &= C_M \left[\prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1}\right] \left(1 - \sum_{j=1}^{M-2} \mu_j\right)^{\alpha_{M-1}+\alpha_M - 1} \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)} \end{aligned} \tag{72}$$

where we have used (2.265). The right hand side of (72) 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

$$\begin{aligned} C_M &= \frac{\Gamma(\alpha_1 + \ldots + \alpha_M)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_{M-2})\Gamma(\alpha_{M-1}+\alpha_M)} \cdot \frac{\Gamma(\alpha_{M-1}+\alpha_M)}{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)} \\ &= \frac{\Gamma(\alpha_1 + \ldots + \alpha_M)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_M)} \end{aligned} \tag{73}$$

as required.

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

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

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

$$E[\ln \mu_j] = \psi(\alpha_k) - \psi(\alpha_0).$$

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

$$\begin{aligned} \widetilde{\mathrm{H}}[p] = &- \int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \lambda \left(\int p(\mathbf{x}) \, \mathrm{d}\mathbf{x} - 1 \right) \\ &+ \mathbf{m}^{\mathrm{T}} \left(\int p(\mathbf{x}) \mathbf{x} \, \mathrm{d}\mathbf{x} - \boldsymbol{\mu} \right) \\ &+ \mathrm{Tr} \left\{ \mathbf{L} \left(\int p(\mathbf{x}) (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \, \mathrm{d}\mathbf{x} - \boldsymbol{\Sigma} \right) \right\}. \end{aligned} \tag{74}$$

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

$$0 = -1 - \ln p(\mathbf{x}) + \lambda + \mathbf{m}^{\mathrm{T}} \mathbf{x} + \mathrm{Tr}\{\mathbf{L}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\}. \tag{75}$$

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

$$p(\mathbf{x}) = \exp\left\{ \lambda - 1 + \mathbf{m}^{\mathrm{T}} \mathbf{x} + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{L} (\mathbf{x} - \boldsymbol{\mu}) \right\}. \tag{76}$$

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

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

2.16 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

$$p(x) = \left(\frac{\tau_1}{2\pi}\right)^{1/2}\left(\frac{\tau_2}{2\pi}\right)^{1/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{\tau_1}{2}(x - \mu_1 - x_2)^2 - \frac{\tau_2}{2}(x_2 - \mu_2)^2\right\}\mathrm{d}x_2. \tag{77}$$

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 (77) which depend on x_2 which are given by

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

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 (77). 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

$$H(x)=\frac{1}{2}\ln\left\{\frac{2\pi(\tau_1+\tau_2)}{\tau_1\tau_2}\right\}. \tag{78}$$

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

$$\Lambda_{ij}^{\mathrm{S}}=\frac{\Lambda_{ij}+\Lambda_{ji}}{2}, \qquad \Lambda_{ij}^{\mathrm{A}}=\frac{\Lambda_{ij}-\Lambda_{ji}}{2} \tag{79}$$

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

$$\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(x_i-\mu_i)\Lambda_{ij}(x_j-\mu_j) \tag{80}$$

where $\mathbf{\Lambda}=\mathbf{\Sigma}^{-1}$ is the precision matrix. When we substitute $\mathbf{\Lambda}=\mathbf{\Lambda}^{\mathrm{A}}+\mathbf{\Lambda}^{\mathrm{S}}$ into (80) 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.

2.20 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.

2.22 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.

2.24 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

$$\mathbf{A}\mathbf{M} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{I} \tag{81}$$

upper right

$$\begin{aligned} &-\mathbf{A}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} + \mathbf{B}\mathbf{D}^{-1} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ &= -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} + \mathbf{B}\mathbf{D}^{-1} \\ &= -\mathbf{B}\mathbf{D}^{-1} + \mathbf{B}\mathbf{D}^{-1} = \mathbf{0} \end{aligned} \tag{82}$$

lower left

$$\mathbf{CM} - \mathbf{DD}^{-1}\mathbf{CM} = \mathbf{CM} - \mathbf{CM} = \mathbf{0} \tag{83}$$

lower right

$$-\mathbf{CMBD}^{-1} + \mathbf{DD}^{-1} + \mathbf{DD}^{-1}\mathbf{CMBD}^{-1} = \mathbf{DD}^{-1} = \mathbf{I}. \tag{84}$$

Thus the right hand side also equals the identity matrix.

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

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

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.

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

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}). \tag{85}$$

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

$$\begin{aligned} &= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x} + \mathbf{x}^{\mathrm{T}}\left[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b})\right] + \mathrm{const} \\ &= -\frac{1}{2}(\mathbf{x} - \mathbf{m})^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})(\mathbf{x} - \mathbf{m}) \\ &\quad + \frac{1}{2}\mathbf{m}^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{m} + \mathrm{const} \end{aligned} \tag{86}$$

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 (86). Then we extract the terms in $\mathbf{y}$ from the final term in (86) and combine these with the remaining terms from the quadratic form (85) which depend on $\mathbf{y}$ to give

$$\begin{aligned} &= -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\left\{\mathbf{L} - \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\right\}\mathbf{y} \\ &\quad + \mathbf{y}^{\mathrm{T}}\left[\left\{\mathbf{L} - \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\right\}\mathbf{b}\right. \\ &\quad \left. + \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\right]. \end{aligned} \tag{87}$$

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

$$\left\{\mathbf{L}-\mathbf{L}\mathbf{A}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\right\}^{-1}=\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \tag{88}$$

which corresponds to (2.110).

Next we identify the mean $\boldsymbol{\nu}$ of the marginal distribution. To do this we make use of (88) in (87) 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).

2.34 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.

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

$$\begin{aligned}
\sigma^2_{(N)} &= \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2 \\
&= \frac{1}{N} \sum_{n=1}^{N-1} (x_n - \mu)^2 + \frac{(x_N - \mu)^2}{N} \\
&= \frac{N-1}{N} \sigma^2_{(N-1)} + \frac{(x_N - \mu)^2}{N} \\
&= \sigma^2_{(N-1)} - \frac{1}{N} \sigma^2_{(N-1)} + \frac{(x_N - \mu)^2}{N} \\
&= \sigma^2_{(N-1)} + \frac{1}{N} \left\{(x_N - \mu)^2 - \sigma^2_{(N-1)}\right\}. \qquad (89)
\end{aligned}$$

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

$$\begin{aligned}\sigma^2_{(N)} &= \sigma^2_{(N-1)} + a_{N-1}\frac{\partial}{\partial\sigma^2_{(N-1)}}\left\{-\frac{1}{2}\ln\sigma^2_{(N-1)} - \frac{(x_N-\mu)^2}{2\sigma^2_{(N-1)}}\right\} \\ &= \sigma^2_{(N-1)} + a_{N-1}\left\{-\frac{1}{2\sigma^2_{(N-1)}} + \frac{(x_N-\mu)^2}{2\sigma^4_{(N-1)}}\right\} \\ &= \sigma^2_{(N-1)} + \frac{a_{N-1}}{2\sigma^4_{(N-1)}}\left\{(x_N-\mu)^2 - \sigma^2_{(N-1)}\right\}. \end{aligned} \tag{90}$$

Comparison of (90) with (89) allows us to identify

$$a_{N-1} = \frac{2\sigma^4_{(N-1)}}{N}. \tag{91}$$

Note that the sign in (2.129) is incorrect, and this equation should read

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

Also, 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 in (2.133). Finally, the labels μ and μ_{ML} in Figure 2.11 should be interchanged.

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

$$p(\boldsymbol{\mu}|\mathbf{X}) \propto p(\boldsymbol{\mu})\prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu},\boldsymbol{\Sigma}). \tag{92}$$

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

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

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

$$\boldsymbol{\mu}_N = \left(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}N\boldsymbol{\mu}_{\mathrm{ML}}\right) \tag{93}$$

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1} \tag{94}$$

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

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n. \tag{95}$$

2.46 From (2.158), we have

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

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

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

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:

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

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

$$\begin{aligned}\mathrm{St}(x|\mu,\lambda,\nu) &\propto \left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-(\nu-1)/2}\\ &= \exp\left(-\frac{\nu-1}{2}\ln\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]\right).\end{aligned} \tag{96}$$

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

$$\ln(1+\epsilon) = \epsilon + O(\epsilon^2) \tag{97}$$

to re-write (96) 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.

2.51 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

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

2.56 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

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\left\{(a-1)\ln\mu + (b-1)\ln(1-\mu)\right\} \tag{98}$$

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

$$h(\mu) = 1 \tag{99}$$

$$g(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{100}$$

$$\mathbf{u}(\mu) = \begin{pmatrix} \ln \mu \\ \ln(1-\mu) \end{pmatrix} \tag{101}$$

$$\boldsymbol{\eta}(a,b) = \begin{pmatrix} a-1 \\ b-1 \end{pmatrix}. \tag{102}$$

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

$$h(\lambda) = 1 \tag{103}$$

$$g(a,b) = \frac{b^a}{\Gamma(a)} \tag{104}$$

$$\mathbf{u}(\lambda) = \begin{pmatrix} \lambda \\ \ln\lambda \end{pmatrix} \tag{105}$$

$$\boldsymbol{\eta}(a,b) = \begin{pmatrix} -b \\ a-1 \end{pmatrix}. \tag{106}$$

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

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

from which we find

$$h(\theta) = 1 \tag{107}$$

$$g(\theta_0, m) = \frac{1}{2\pi I_0(m)} \tag{108}$$

$$\mathbf{u}(\theta) = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{109}$$

$$\boldsymbol{\eta}(\theta_0, m) = \begin{pmatrix} m\cos\theta_0 \\ m\sin\theta_0 \end{pmatrix}. \tag{110}$$

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.

Chapter 3 Linear Models for Regression

3.1 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 $u_0 = w_0 + \sum_{j=1}^{M} w_j/2$. Note that there is a typographical error in the question: there is a 2 missing in the denominator of the argument to the '$\tanh$' function in equation (3.102).

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.

3.8 Combining the prior

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

and the likelihood

$$p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left(-\frac{\beta}{2}(t_{N+1} - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2\right) \tag{111}$$

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

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

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

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

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

$$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}. \tag{112}$$

and

$$\mathbf{m}_{N+1} = \mathbf{S}_{N+1}(\mathbf{S}_N^{-1}\mathbf{m}_N + \beta\boldsymbol{\phi}_{N+1}t_{N+1}). \tag{113}$$

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).

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

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

3.18 We can rewrite (3.79)

$$\begin{aligned} &\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &= \frac{\beta}{2}\left(\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w}\right) + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right) \end{aligned}$$

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:

$$\begin{aligned} &\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right) \\ &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{A}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right) \\ &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} + \mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N - \mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right) \\ &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right) + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N). \end{aligned}$$

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.

3.20 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.

3.23 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

$$
\begin{aligned}
p(\mathbf{t}) &= \frac{b_0^{a_0}}{\left((2\pi)^{M+N}|\mathbf{S}_0|\right)^{1/2}} (2\pi)^{M/2}|\mathbf{S}_N|^{1/2} \int \beta^{a_N-1}\exp(-b_N\beta)\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{N/2}} \frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{\Gamma(a_N)}{\Gamma(a_0)}.
\end{aligned}
$$

Chapter 4 Linear Models for Classification

4.2 For the purpose of this exercise, we make the contribution of the bias weights explicit in (4.15), giving

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})\right\}, \tag{114}$$

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 (114) 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

$$\mathbf{w}_0 = \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \tag{115}$$

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 (115) into (114), 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 $\mathbf{x}^\star$,

$$\begin{aligned} \mathbf{y}(\mathbf{x}^\star) &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \mathbf{w}_0 \\ &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \\ &= \bar{\mathbf{t}} - \widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}(\mathbf{x}^\star - \bar{\mathbf{x}}). \end{aligned} \tag{116}$$

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 (116), 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 $\mathbf{a}^{\mathrm{T}}\widehat{\mathbf{T}}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}(\mathbf{T} - \overline{\mathbf{T}})^{\mathrm{T}} = b(\mathbf{1} - \mathbf{1})^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$.

4.4 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

$$\nabla L = \mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} \tag{117}$$

and setting this gradient to zero gives

$$\mathbf{w} = -\frac{1}{2\lambda}(\mathbf{m}_2 - \mathbf{m}_1)$$

form which it follows that $\mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1$.

4.7 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

$$\begin{aligned}y = \sigma(a) &= \frac{1}{1+e^{-a}} \\ \Rightarrow \quad \frac{1}{y} - 1 &= e^{-a} \\ \Rightarrow \quad \ln\left\{\frac{1-y}{y}\right\} &= -a \\ \Rightarrow \quad \ln\left\{\frac{y}{1-y}\right\} &= a = \sigma^{-1}(y).\end{aligned}$$

4.9 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

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln p(\phi_n|\mathcal{C}_k) + \ln \pi_k\right\}. \tag{118}$$

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

$$-\pi_k\lambda = \sum_{n}^{N} t_{nk} = N_k. \tag{119}$$

Summing both sides over k we find that $\lambda = -N$, and using this to eliminate λ we obtain (4.159).

4.12 Differentiating (4.59) we obtain

$$\begin{aligned}
\frac{d\sigma}{da} &= \frac{e^{-a}}{\left(1+e^{-a}\right)^2} \\
&= \sigma(a)\left\{\frac{e^{-a}}{1+e^{-a}}\right\} \\
&= \sigma(a)\left\{\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right\} \\
&= \sigma(a)(1-\sigma(a)).
\end{aligned}$$

4.13 We start by computing the derivative of (4.90) w.r.t. y_n

$$\begin{aligned}
\frac{\partial E}{\partial y_n} &= \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} && (120)\\
&= \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)} \\
&= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} && (121)\\
&= \frac{y_n - t_n}{y_n(1-y_n)}. && (122)
\end{aligned}$$

From (4.88), we see that

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)\left(1 - \sigma(a_n)\right) = y_n(1 - y_n). \tag{123}$$

Finally, we have

$$\nabla a_n = \boldsymbol{\phi}_n \tag{124}$$

where ∇ denotes the gradient with respect to $\mathbf{w}$. Combining (122), (123) and (124) 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.

4.17 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).

4.19 Using the cross-entropy error function (4.90), and following Exercise 4.13, we have

$$\frac{\partial E}{\partial y_n} = \frac{y_n - t_n}{y_n(1 - y_n)}. \tag{125}$$

Also

$$\nabla a_n = \boldsymbol{\phi}_n. \tag{126}$$

From (4.115) and (4.116) we have

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \Phi(a_n)}{\partial a_n} = \frac{1}{\sqrt{2\pi}} e^{-a_n^2}. \tag{127}$$

Combining (125), (126) and (127), we get

$$\nabla E = \sum_{n=1}^{N} \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial a_n}\nabla a_n = \sum_{n=1}^{N} \frac{y_n - t_n}{y_n(1 - y_n)}\frac{1}{\sqrt{2\pi}} e^{-a_n^2}\boldsymbol{\phi}_n. \tag{128}$$

In order to find the expression for the Hessian, it is is convenient to first determine

$$\begin{aligned}\frac{\partial}{\partial y_n}\frac{y_n - t_n}{y_n(1-y_n)} &= \frac{y_n(1-y_n)}{y_n^2(1-y_n)^2} - \frac{(y_n - t_n)(1-2y_n)}{y_n^2(1-y_n)^2} \\ &= \frac{y_n^2 + t_n - 2y_n t_n}{y_n^2(1-y_n)^2}. \end{aligned} \tag{129}$$

Then using (126)–(129) we have

$$\begin{aligned}\nabla\nabla E &= \sum_{n=1}^{N}\left\{\frac{\partial}{\partial y_n}\left[\frac{y_n - t_n}{y_n(1-y_n)}\right]\frac{1}{\sqrt{2\pi}}e^{-a_n^2}\boldsymbol{\phi}_n\nabla y_n \right. \\ &\quad \left. + \frac{y_n - t_n}{y_n(1-y_n)}\frac{1}{\sqrt{2\pi}}e^{-a_n^2}(-2a_n)\boldsymbol{\phi}_n\nabla a_n\right\} \\ &= \sum_{n=1}^{N}\left(\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)}\frac{1}{\sqrt{2\pi}}e^{-a_n^2} - 2a_n(y_n - t_n)\right)\frac{e^{-2a_n^2}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}}{\sqrt{2\pi}y_n(1-y_n)}.\end{aligned}$$

4.23 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

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const} \tag{130}$$

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 (130) 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 (130), we obtain (4.139) by dropping the $\ln|\widehat{\mathbf{H}}|$ since this $O(1)$ compared to $\ln N$.

Chapter 5 Neural Networks

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

$$\begin{aligned}
&\sum_{n=1}^{N}\ln\mathcal{N}\left(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n,\mathbf{w}),\beta^{-1}\mathbf{I}\right)\\
&\quad= -\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{t}_n-\mathbf{y}(\mathbf{x}_n,\mathbf{w})\right)^{\mathrm{T}}(\beta\mathbf{I})\left(\mathbf{t}_n-\mathbf{y}(\mathbf{x}_n,\mathbf{w})\right)+\text{const}\\
&\quad= -\frac{\beta}{2}\sum_{n=1}^{N}\|\mathbf{t}_n-\mathbf{y}(\mathbf{x}_n,\mathbf{w})\|^2+\text{const},
\end{aligned}$$

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.

5.5 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

$$p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}.$$

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

$$\frac{\partial E}{\partial a_n} = -t_n \frac{1}{y_n}\frac{\partial y_n}{\partial a_n} + (1-t_n)\frac{1}{1-y_n}\frac{\partial y_n}{\partial a_n}. \tag{131}$$

From (4.88), we have

$$\frac{\partial y_n}{\partial a_n} = y_n(1-y_n). \tag{132}$$

Substituting (132) into (131), 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.

5.9 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

$$\begin{aligned}E(\mathbf{w}) &= -\sum_{n}^{N}\frac{1+t_n}{2}\ln\frac{1+y_n}{2} + \left(1-\frac{1+t_n}{2}\right)\ln\left(1-\frac{1+y_n}{2}\right) \\ &= -\frac{1}{2}\sum_{n}^{N}\{(1+t_n)\ln(1+y_n) + (1-t_n)\ln(1-y_n)\} + N\ln 2\end{aligned}$$

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

$$\lambda_i = \mathbf{u}_i^{\mathrm{T}}\mathbf{H}\mathbf{u}_i > 0 \tag{133}$$

for all values of i. Thus, if $\mathbf{H}$ is positive definite, all of its eigenvalues will be positive.

Conversely, assume that (133) 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 (133). Thus, if all of the eigenvalues are positive, the Hessian matrix will be positive definite.

5.11 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 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.

5.19 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 $\mathbf{b}_n \equiv \nabla a_n$.

5.25 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

$$\begin{aligned} w_j^{(\tau)} &= \mathbf{u}_j^{\mathrm{T}} \mathbf{w}^{(\tau)} && (134) \\ &= \mathbf{u}_j^{\mathrm{T}} \mathbf{w}^{(\tau-1)} - \rho \mathbf{u}_j^{\mathrm{T}} \mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^\star) \\ &= w_j^{(\tau-1)} - \rho \eta_j \mathbf{u}_j^{\mathrm{T}} (\mathbf{w} - \mathbf{w}^\star) \\ &= w_j^{(\tau-1)} - \rho \eta_j (w_j^{(\tau-1)} - w_j^\star), && (135) \end{aligned}$$

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 (135), 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 (135)

$$\begin{aligned} w_j^{(N)} &= w_j^{(N-1)} - \rho\eta_j(w_j^{(N-1)} - w_j^\star) \\ &= w_j^{(N-1)}(1 - \rho\eta_j) + \rho\eta_j w_j^\star \\ &= \left\{1 - (1 - \rho\eta_j)^{N-1}\right\} w_j^\star(1 - \rho\eta_j) + \rho\eta_j w_j^\star \\ &= \left\{(1 - \rho\eta_j) - (1 - \rho\eta_j)^N\right\} w_j^\star + \rho\eta_j w_j^\star \\ &= \left\{1 - (1 - \rho\eta_j)^N\right\} w_j^\star \end{aligned}$$

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

$$\begin{aligned} |w_j^{(\tau)}| &= |\{1 - (1 - \rho\eta_j)^\tau\} w_j^\star| \\ &= |\left\{1 - (1 - \tau\rho\eta_j + O(\rho\eta_j^2))\right\} w_j^\star| \\ &\simeq \tau\rho\eta_j|w_j^\star| \ll |w_j^\star| \end{aligned}$$

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 $\rho\tau$.

5.27 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).

5.34 We start by using the chain rule to write

$$\frac{\partial E_n}{\partial a_k^\pi} = \sum_{j=1}^{K} \frac{\partial E_n}{\partial \pi_j} \frac{\partial \pi_j}{\partial a_k^\pi}. \tag{136}$$

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 (136), standard derivatives applied to the n^{th} term of (5.153) gives

$$\frac{\partial E_n}{\partial \pi_j} = -\frac{\mathcal{N}_{nj}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}} = -\frac{\gamma_{nj}}{\pi_j}. \tag{137}$$

For the for the second factor, we have from (4.106) that

$$\frac{\partial \pi_j}{\partial a_k^\pi} = \pi_j (I_{jk} - \pi_k). \tag{138}$$

Combining (136), (137) and (138), 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.

5.39 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), since $p(\mathcal{D}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)$ 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

$$\prod_{n}^{N}\prod_{k}^{K}y_k(\mathbf{x}_n,\mathbf{w})^{t_{nk}}$$

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.

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 $\mathbf{K}=\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$ 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 $\mathbf{Ka}_{\perp}=\mathbf{0}$. 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 $\mathbf{a} = \mathbf{a}_{||}$ where $\mathbf{a}_{||}$ 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

$$\mathbf{a} = \mathbf{\Phi}\mathbf{u}. \tag{139}$$

Substituting (139) into (6.7) we obtain

$$\begin{aligned} J(\mathbf{u}) &= \frac{1}{2}\left(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}\mathbf{u} \\ &= \frac{1}{2}\left(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} \end{aligned} \tag{140}$$

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 (140) we recover the original regularized error function (6.2).

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.

6.7 (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.

6.12 **NOTE**: In the first 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.

6.14 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

$$\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) = \nabla_{\mu} \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}) = \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

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.

6.17 **NOTE**: In the first 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

$$y(\mathbf{x}) \rightarrow y(\mathbf{x}) + \epsilon\eta(\mathbf{x}).$$

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

$$\sum_{n=1}^{N}\int \{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\}\,\eta(\mathbf{x}_n + \boldsymbol{\xi})\nu(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = 0. \tag{141}$$

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

$$\sum_{n=1}^{N} \int \{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\}\, \delta(\mathbf{x}_n + \boldsymbol{\xi} - \mathbf{z}) \nu(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi} = \sum_{n=1}^{N} \{y(\mathbf{z}) - t_n\}\, \nu(\mathbf{z} - \mathbf{x}_n).$$

Substing this back into (141) and rearranging we then obtain the required result (6.40).

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

$$\mathbf{C}_N = \frac{1}{\alpha}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N \tag{142}$$

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 (142), (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 (142) 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

$$\begin{aligned}\sigma^2_{N+1}(\mathbf{x}_{N+1}) &= \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_{N+1}) + \beta^{-1} \\ &\quad -\alpha^{-2}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\left(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}_{N+1}) \\ &= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\big(\alpha^{-1}\mathbf{I}_M \\ &\quad -\alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\left(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\big)\boldsymbol{\phi}(\mathbf{x}_{N+1}). \qquad (143)\end{aligned}$$

We now make use of the matrix identity (C.7) to give

$$\begin{aligned}\alpha^{-1}\mathbf{I}_M - \alpha^{-1}\mathbf{I}_M\boldsymbol{\Phi}^{\mathrm{T}}\left(\boldsymbol{\Phi}(\alpha^{-1}\mathbf{I}_M)\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\alpha^{-1}\mathbf{I}_M \\ = \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1} = \mathbf{S}_N,\end{aligned}$$

where we have also used (3.54). Substituting this in (143), 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.

6.23 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.

6.25 Substituting the gradient and the Hessian into the Newton-Raphson formula we obtain

$$\begin{aligned}\mathbf{a}_N^{\text{new}} &= \mathbf{a}_N + (\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1}\mathbf{a}_N\right] \\ &= (\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N\mathbf{a}_N\right] \\ &= \mathbf{C}_N(\mathbf{I} + \mathbf{W}_N\mathbf{C}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N\mathbf{a}_N\right]\end{aligned}$$

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 $\phi(\mathbf{x})$.

7.4 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

$$\frac{1}{2}\|\mathbf{w}\|^2 = \sum_{n}^{N} a_n - \frac{1}{2}\|\mathbf{w}\|^2,$$

from which the desired result follows.

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

$$\begin{aligned} \xi_n &\geqslant 0 \\ \mu_n &\geqslant 0 \\ \mu_n\xi_n &= 0 \end{aligned} \tag{144}$$

and from (7.59) we have that

$$\mu_n = C - a_n. \tag{145}$$

Combining (144) and (145), we get (7.67); similar reasoning for $\widehat{\mu}_n$ and $\widehat{\xi}_n$ lead to (7.68).

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)

$$p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \frac{1}{(2\pi)^{N/2}} \prod_{i=1}^{M} \alpha_i \int \exp\{-E(\mathbf{w})\}\, \mathrm{d}\mathbf{w} \tag{146}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}$$

and $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$.

Completing the square over $\mathbf{w}$, we get

$$E(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{m})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \mathbf{m}) + E(\mathbf{t}) \tag{147}$$

where $\mathbf{m}$ and $\boldsymbol{\Sigma}$ are given by (7.82) and (7.83), respectively, and

$$E(\mathbf{t}) = \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m}\right). \tag{148}$$

Using (147), we can evaluate the integral in (146) to obtain

$$\int \exp\{-E(\mathbf{w})\}\, \mathrm{d}\mathbf{w} = \exp\{-E(\mathbf{t})\}\, (2\pi)^{M/2} |\boldsymbol{\Sigma}|^{1/2}. \tag{149}$$

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 (148) as follows

$$\begin{aligned}
E(\mathbf{t}) &= \frac{1}{2}\left(\beta \mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m}\right) \\
&= \frac{1}{2}\left(\beta \mathbf{t}^{\mathrm{T}}\mathbf{t} - \beta \mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\beta\right) \\
&= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta \mathbf{I} - \beta\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\beta\right)\mathbf{t} \\
&= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta \mathbf{I} - \beta\boldsymbol{\Phi}(\mathbf{A} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\beta\right)\mathbf{t} \\
&= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\right)^{-1}\mathbf{t} \\
&= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t}.
\end{aligned}$$

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 $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta)$.

7.12 Using the results (146)–(149) from Solution 7.10, we can write (7.85) in the form of (3.86):

$$\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{N}{2}\ln\beta + \frac{1}{2}\sum_{i}^{N}\ln\alpha_i - E(\mathbf{t}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\ln(2\pi). \tag{150}$$

By making use of (148) and (7.83) together with (C.22), we can take the derivatives of this w.r.t α_i, yielding

$$\frac{\partial}{\partial\alpha_i}\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2}m_i^2. \tag{151}$$

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

$$\frac{\partial}{\partial\beta}\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{1}{2}\left(\frac{N}{\beta} - \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 - \mathrm{Tr}\left[\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right]\right). \tag{152}$$

Using (7.83), we can rewrite the argument of the trace operator as

$$\begin{aligned}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} &= \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\ &= \boldsymbol{\Sigma}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\ &= (\mathbf{A} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\ &= (\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma})\beta^{-1}. \end{aligned} \tag{153}$$

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 (152) equal to zero and then re-arrange to obtain (7.88).

7.15 Using (7.94), (7.95) and (7.97)–(7.99), we can rewrite (7.85) as follows

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) &= -\frac{1}{2}\Bigg\{N\ln(2\pi) + \ln|\mathbf{C}_{-i}||1 + \alpha_i^{-1}\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| \\ &\quad + \mathbf{t}^{\mathrm{T}}\left(\mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}\right)\mathbf{t}\Bigg\} \\ &= -\frac{1}{2}\left\{N\ln(2\pi) + \ln|\mathbf{C}_{-i}| + \mathbf{t}^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\mathbf{t}\right\} \\ &\quad + \frac{1}{2}\left[-\ln|1 + \alpha_i^{-1}\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| + \mathbf{t}^{\mathrm{T}}\frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}\mathbf{t}\right] \\ &= L(\alpha_{-i}) + \frac{1}{2}\left[\ln\alpha_i - \ln(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right] \\ &= L(\alpha_{-i}) + \lambda(\alpha_i)\end{aligned}$$

7.18 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 (123) and (124) 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.

Chapter 8 Probabilistic Graphical Models

CHECK! 8.1 We want to show that, for (8.5),

$$\sum_{x_1}\dots\sum_{x_K} p(\mathbf{x}) = \sum_{x_1}\dots\sum_{x_K}\prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) = 1.$$

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

$$\begin{aligned}\sum_{x_1}\dots\sum_{x_K} p(\mathbf{x}) &= \sum_{x_1}\dots\sum_{x_K} p(x_K|\mathrm{pa}_K)\prod_{k=1}^{K-1} p(x_k|\mathrm{pa}_k)\\ &= \sum_{x_1}\dots\sum_{x_{K-1}}\prod_{k=1}^{K-1} p(x_k|\mathrm{pa}_k),\end{aligned}$$

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.

8.5 The solution is given in Figure 3.

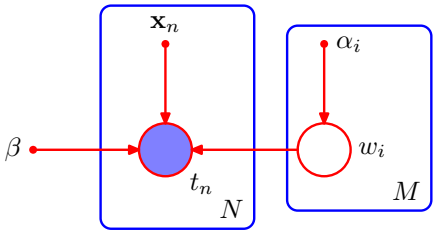


Figure 3 The graphical representation of the relevance vector machine (RVM); Solution 8.5.

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.

8.12 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 4.

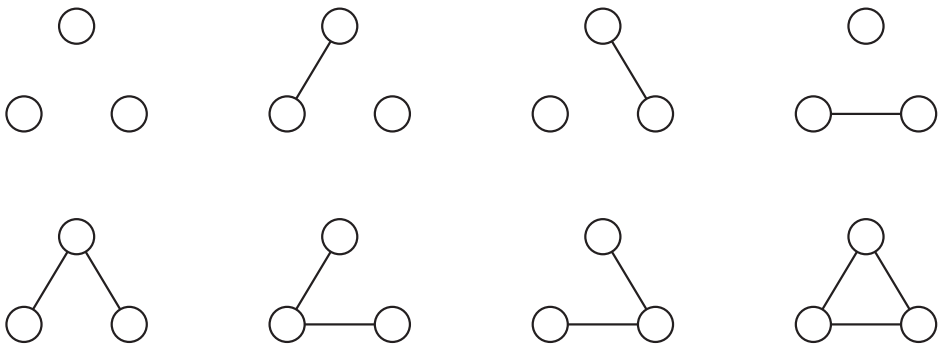


Figure 4 The set of 8 distinct undirected graphs which can be constructed over $M = 3$ nodes.

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),

$$p(x_n) = \frac{1}{Z} \underbrace{\left[\sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \cdots \left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right] \cdots\right]}_{\mu_\alpha(x_{n-1})} \psi_{n-1,n}(x_{n-1}, x_n)$$
$$\underbrace{\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \cdots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots\right]}_{\mu_\beta(x_n)},$$

from which (8.58) immediately follows.

8.18 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 5. 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.

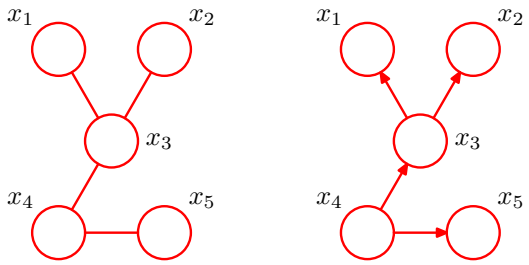


Figure 5 The graph on the left is an undirected tree. If we pick x_4 to be the root node and direct all the edges in the graph to point from the root to the leaf nodes (x_1, x_2 and x_5), we obtain the directed tree shown on the right.

Since 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.

8.20 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 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.

8.23 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

$$\begin{aligned} p(x_i) &= \prod_{s \in \mathrm{ne}(x_i)} \mu_{f_s \to x_i}(x_i) \\ &= \mu_{f_s \to x_i}(x_i) \prod_{t \in \mathrm{ne}(x_i) \setminus f_s} \mu_{f_t \to x_i}(x_i) \\ &= \mu_{f_s \to x_i}(x_i)\, \mu_{x_i \to f_s}(x_i). \end{aligned}$$

8.28 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

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 $\{\boldsymbol{\mu}_k\}$.

9.3 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.

9.7 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).

9.12 Since the expectation of a sum is the sum of the expectations we have

$$\mathrm{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \mathrm{E}_k[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$$

where $\mathrm{E}_k[\mathbf{x}]$ denotes the expectation of $\mathbf{x}$ under the distribution $p(\mathbf{x}|k)$. To find the covariance we use the general relation

$$\mathrm{cov}[\mathbf{x}] = \mathrm{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathrm{E}[\mathbf{x}]\mathrm{E}[\mathbf{x}]^{\mathrm{T}}$$

to give

$$\begin{aligned}
\mathrm{cov}[\mathbf{x}] &= \mathrm{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathrm{E}[\mathbf{x}]\mathrm{E}[\mathbf{x}]^{\mathrm{T}} \\
&= \sum_{k=1}^{K} \pi_k \mathrm{E}_k[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathrm{E}[\mathbf{x}]\mathrm{E}[\mathbf{x}]^{\mathrm{T}} \\
&= \sum_{k=1}^{K} \pi_k \left\{ \boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathrm{T}} \right\} - \mathrm{E}[\mathbf{x}]\mathrm{E}[\mathbf{x}]^{\mathrm{T}}.
\end{aligned}$$

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.

9.17 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.

9.20 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).

9.23 **NOTE**: In the first 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 the case when the optimization has converged.

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 $\alpha_i^\star=\alpha_i^{\mathrm{new}}=\alpha_i$ 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

$$\frac{1}{\beta^\star}=\frac{\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}_N\|^2}{N}+\frac{\sum_i\gamma_i}{\beta^\star N}.$$

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).

9.25 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

$$N_k^{\text{old}} = \sum_n \gamma^{\text{old}}(z_{nk}). \tag{154}$$

We get N_k^{new} by recomputing the responsibilities, $\gamma(z_{mk})$, for a specific data point, $\mathbf{x}_m$, yielding

$$N_k^{\text{new}} = \sum_{n\neq m} \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}).$$

Combining this with (154), 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

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}}\left(\sum_{n\neq m} \gamma^{\text{old}}(z_{nk})\mathbf{x}_n + \gamma^{\text{new}}(z_{mk})\mathbf{x}_m\right) \\
&= \frac{1}{N_k^{\text{new}}}\left(N_k^{\text{old}}\boldsymbol{\mu}_k^{\text{old}} - \gamma^{\text{old}}(z_{mk})\mathbf{x}_m + \gamma^{\text{new}}(z_{mk})\mathbf{x}_m\right) \\
&= \frac{1}{N_k^{\text{new}}}\Big(\left(N_k^{\text{new}} - \gamma^{\text{new}}(z_{mk}) + \gamma^{\text{old}}(z_{mk})\right)\boldsymbol{\mu}_k^{\text{old}} \\
&\qquad -\gamma^{\text{old}}(z_{mk})\mathbf{x}_m + \gamma^{\text{new}}(z_{mk})\mathbf{x}_m\Big) \\
&= \boldsymbol{\mu}_k^{\text{old}} + \left(\frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}}\right)(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}),
\end{aligned}$$

where we have used (9.79).

Chapter 10 Variational Inference and EM

10.1 Starting from (10.3), we use the product rule together with (10.4) to get

$$\begin{aligned}\mathcal{L}(q) &= \int q\left(\mathbf{Z}\right) \ln\left\{\frac{p\left(\mathbf{X},\mathbf{Z}\right)}{q\left(\mathbf{Z}\right)}\right\} \mathrm{d}\mathbf{Z} \\ &= \int q\left(\mathbf{Z}\right) \ln\left\{\frac{p\left(\mathbf{X} \mid \mathbf{Z}\right)\, p\left(\mathbf{X}\right)}{q\left(\mathbf{Z}\right)}\right\} \mathrm{d}\mathbf{Z} \\ &= \int q\left(\mathbf{Z}\right) \left(\ln\left\{\frac{p\left(\mathbf{X} \mid \mathbf{Z}\right)}{q\left(\mathbf{Z}\right)}\right\} + \ln p\left(\mathbf{X}\right)\right) \mathrm{d}\mathbf{Z} \\ &= -\mathrm{KL}(\, q \parallel p\,) + \ln p\left(\mathbf{X}\right).\end{aligned}$$

Rearranging this, we immediately get (10.2).

10.3 Starting from (10.16) and optimizing w.r.t. $q_j\left(\mathbf{Z}_j\right)$, we get

$$\begin{aligned}\mathrm{KL}(\, p \parallel q\,) &= -\int p\left(\mathbf{Z}\right) \left[\sum_{i=1}^{M} \ln q_i\left(\mathbf{Z}_i\right)\right] \mathrm{d}\mathbf{Z} + \text{const.} \\ &= -\int \left(p\left(\mathbf{Z}\right) \ln q_j\left(\mathbf{Z}_j\right) + p\left(\mathbf{Z}\right) \sum_{i \neq j} \ln q_i\left(\mathbf{Z}_i\right)\right) \mathrm{d}\mathbf{Z} + \text{const.} \\ &= -\int p\left(\mathbf{Z}\right) \ln q_j\left(\mathbf{Z}_j\right)\, \mathrm{d}\mathbf{Z} + \text{const.} \\ &= -\int \ln q_j\left(\mathbf{Z}_j\right) \left[\int p\left(\mathbf{Z}\right) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i\right] \mathrm{d}\mathbf{Z}_j + \text{const.} \\ &= -\int F_j(\mathbf{Z}_j) \ln q_j\left(\mathbf{Z}_j\right)\, \mathrm{d}\mathbf{Z}_j + \text{const.},\end{aligned}$$

where terms independent of $q_j\left(\mathbf{Z}_j\right)$ have been absorbed into the constant term and we have defined

$$F_j(\mathbf{Z}_j) = \int p\left(\mathbf{Z}\right) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i.$$

We use a Lagrange multiplier to ensure that $q_j\left(\mathbf{Z}_j\right)$ integrates to one, yielding

$$-\int F_j(\mathbf{Z}_j) \ln q_j\left(\mathbf{Z}_j\right)\, \mathrm{d}\mathbf{Z}_j + \lambda\left(\int q_j\left(\mathbf{Z}_j\right)\, \mathrm{d}\mathbf{Z}_j - 1\right).$$

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

$$-\frac{F_j(\mathbf{Z}_j)}{q_j\left(\mathbf{Z}_j\right)} + \lambda = 0.$$

From this, we see that

$$\lambda q_j\left(\mathbf{Z}_j\right) = F_j(\mathbf{Z}_j).$$

Integrating both sides over $\mathbf{Z}_j$, we see that, since $q_j\left(\mathbf{Z}_j\right)$ must intgrate to one,

$$\lambda = \int F_j(\mathbf{Z}_j)\,\mathrm{d}\mathbf{Z}_j = \int \left[\int p\left(\mathbf{Z}\right) \prod_{i\neq j} \mathrm{d}\mathbf{Z}_i\right] \mathrm{d}\mathbf{Z}_j = 1,$$

and thus

$$q_j\left(\mathbf{Z}_j\right) = F_j(\mathbf{Z}_j) = \int p\left(\mathbf{Z}\right) \prod_{i\neq j} \mathrm{d}\mathbf{Z}_i.$$

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

$$\mathrm{KL}(\,q \parallel p\,) = -\iint q\left(\boldsymbol{\theta}\right) q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{X}\right)}{q\left(\mathbf{z}\right) q\left(\boldsymbol{\theta}\right)}\,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\theta}.$$

For the particular chosen form of $q(\boldsymbol{\theta})$, this is equivalent to

$$\begin{aligned}
\mathrm{KL}(\,q \parallel p\,) &= -\int q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{z}, \boldsymbol{\theta}_0 \mid \mathbf{X}\right)}{q\left(\mathbf{z}\right)}\,\mathrm{d}\mathbf{z} + \text{const.} \\
&= -\int q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{z} \mid \boldsymbol{\theta}_0, \mathbf{X}\right) p\left(\boldsymbol{\theta}_0 \mid \mathbf{X}\right)}{q\left(\mathbf{z}\right)}\,\mathrm{d}\mathbf{z} + \text{const.} \\
&= -\int q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{z} \mid \boldsymbol{\theta}_0, \mathbf{X}\right)}{q\left(\mathbf{z}\right)}\,\mathrm{d}\mathbf{z} + \text{const.},
\end{aligned}$$

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

$$\begin{aligned}
&\int q\left(\boldsymbol{\theta}\right) \int q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{X}, \boldsymbol{\theta}, \mathbf{z}\right)}{q\left(\boldsymbol{\theta}\right) q\left(\mathbf{z}\right)}\,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\theta} \\
&= \int q\left(\boldsymbol{\theta}\right) \mathbb{E}_{q(\mathbf{z})}\left[\ln p\left(\mathbf{X}, \boldsymbol{\theta}, \mathbf{z}\right)\right]\,\mathrm{d}\boldsymbol{\theta} - \int q\left(\boldsymbol{\theta}\right) \ln q\left(\boldsymbol{\theta}\right)\,\mathrm{d}\boldsymbol{\theta} + \text{const.}
\end{aligned}$$

where the last term summarizes terms independent of $q\left(\boldsymbol{\theta}\right)$. 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

$$\mathbb{E}_{q(\mathbf{z})}\left[\ln p\left(\mathbf{X}, \boldsymbol{\theta}_0, \mathbf{z}\right)\right],$$

w.r.t. $\boldsymbol{\theta}_0$, which is equivalent to the M-step of the EM algorithm.

10.10 **NOTE**: The first printing of PRML contains 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

$$\begin{aligned}\mathcal{L} &= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, \mathbf{X}, m)}{q(\mathbf{Z}|m)q(m)} \right\} \\ &= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \left\{ \ln p(\mathbf{Z}, \mathbf{X}|m) + \ln p(m) - \ln q(\mathbf{Z}|m) - \ln q(m) \right\} \\ &= \sum_m q(m) \Bigg(\ln p(m) - \ln q(m) \\ &\qquad + \sum_{\mathbf{Z}} q(\mathbf{Z}|m) \{ \ln p(\mathbf{Z}, \mathbf{X}|m) - \ln q(\mathbf{Z}|m) \} \Bigg) \\ &= \sum_m q(m) \left\{ \ln \left(p(m) \exp\{\mathcal{L}_m\} \right) - \ln q(m) \right\}, \qquad (155)\end{aligned}$$

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 (155) 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

$$q(m) \propto p(m) \exp\{\mathcal{L}_m\}.$$

10.13 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

$$\begin{aligned}\ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &= \ln \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, \beta_0\boldsymbol{\Lambda}_k) + \ln \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0) \\ &\quad + \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \text{const.} \\ &= -\frac{\beta_0}{2}(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0) + \frac{1}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k\mathbf{W}_0^{-1}\right) \\ &\quad + \frac{(\nu_0 - D - 1)}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) \\ &\quad + \frac{1}{2}\left(\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\right)\ln|\boldsymbol{\Lambda}_k| + \text{const.} \end{aligned} \tag{156}$$

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 (156) 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

$$q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \beta_k\boldsymbol{\Lambda}_k) \tag{157}$$

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 (156), and we substitute for $\ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ using the result (157). Keeping only those terms

which depend on $\boldsymbol{\Lambda}_k$ we obtain

$$\begin{aligned}\ln q^\star(\boldsymbol{\Lambda}_k) &= -\frac{\beta_0}{2}(\boldsymbol{\mu}_k-\mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k-\mathbf{m}_0)+\frac{1}{2}\ln|\boldsymbol{\Lambda}_k|-\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k\mathbf{W}_0^{-1}\right)\\ &\quad+\frac{(\nu_0-D-1)}{2}\ln|\boldsymbol{\Lambda}_k|-\frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)\\ &\quad+\frac{1}{2}\left(\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\right)\ln|\boldsymbol{\Lambda}_k|+\frac{\beta_k}{2}(\boldsymbol{\mu}_k-\mathbf{m}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}-\mathbf{m}_k)\\ &\quad-\frac{1}{2}\ln|\boldsymbol{\Lambda}_k|+\text{const.}\\ &= \frac{(\nu_k-D-1)}{2}\ln|\boldsymbol{\Lambda}_k|-\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k\mathbf{W}_k^{-1}\right)+\text{const.}\end{aligned}$$

Note that the terms involving $\boldsymbol{\mu}_k$ have cancelled out as we expect since $q^\star(\boldsymbol{\Lambda}_k)$ is independent of $\boldsymbol{\mu}_k$. Here we have defined

$$\begin{aligned}\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1}+\beta_0(\boldsymbol{\mu}_k-\mathbf{m}_0)(\boldsymbol{\mu}_k-\mathbf{m}_0)^{\mathrm{T}}+\sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\\ &\quad-\beta_k(\boldsymbol{\mu}_k-\mathbf{m}_k)(\boldsymbol{\mu}_k-\mathbf{m}_k)^{\mathrm{T}}\\ &= \mathbf{W}_0^{-1}+N_k\mathbf{S}_k+\frac{\beta_0 N_k}{\beta_0+N_k}(\overline{\mathbf{x}}_k-\mathbf{m}_0)(\overline{\mathbf{x}}_k-\mathbf{m}_0)^{\mathrm{T}}\\ \nu_k &= \nu_0+\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\\ &= \nu_0+N_k,\end{aligned}$$

where we have made use of the result

$$\begin{aligned}\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} &= \sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n-\overline{\mathbf{x}}_k)(\mathbf{x}_n-\overline{\mathbf{x}}_k)^{\mathrm{T}}+N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}}\\ &= N_k\mathbf{S}_k+N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}}\end{aligned}\tag{158}$$

and we have made use of (10.53). Thus we see that $q^\star(\boldsymbol{\Lambda}_k)$ is a Wishart distribution of the form

$$q^\star(\boldsymbol{\Lambda}_k)=\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k,\nu_k).$$

10.16 To derive (10.71) we make use of (10.38) to give

$$\begin{aligned}&\mathbb{E}[\ln p(D|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})]\\ &= \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left\{\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|]-\mathbb{E}[(\mathbf{x}_n-\boldsymbol{\mu}_k)\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)]-D\ln(2\pi)\right\}.\end{aligned}$$

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

$$\mathbb{E}[\ln p(D|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\big\{\ln \widetilde{\Lambda}_k - D\beta_k^{-1} - \nu_k(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n - \mathbf{m}_k) - D\ln(2\pi)\big\}.$$

Now we use the definitions (10.51) to (10.53) together with the result (158) 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).

10.20 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

$$\begin{aligned} -\ln B(\mathbf{W}_k, \nu_k) &\to -\frac{N_k D}{2}\left(\ln N_k - \ln 2 - \ln N_k + \ln 2 + 1\right) - \frac{N_k}{2}\ln|\mathbf{S}_k| \\ &= -\frac{N_k}{2}\left(\ln|\mathbf{S}_k| + D\right). \end{aligned} \tag{159}$$

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

$$\begin{aligned}\mathbb{E}\left[\ln|\mathbf{\Lambda}|\right] &\to D\ln\frac{N_k}{2} + D\ln 2 - D\ln N_k - \ln|\mathbf{S}_k| \\ &= -\ln|\mathbf{S}_k|,\end{aligned}$$

where we approximated the argument to the digamma function by $N_k/2$. Substituting this and (159) into (B.82), we get

$$\mathrm{H}[\mathbf{\Lambda}] \to 0$$

when $N_k \to \infty$.

Next consider the posterior distribution over the mean $\boldsymbol{\mu}_k$ of the k^{th} component given by

$$q^\star(\boldsymbol{\mu}_k|\mathbf{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \beta_k\mathbf{\Lambda}_k).$$

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.

10.23 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

$$\frac{N_k}{\pi_k} + \lambda = 0, \tag{160}$$

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 (160) 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

$$(\nu_0 + N_k - D)\ln|\boldsymbol{\Lambda}_k| - \mathrm{Tr}\left[(\mathbf{W}_0^{-1} + \beta_0(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + N_k\mathbf{S}_k)\boldsymbol{\Lambda}_k\right],$$

where we have dropped the constant term. Using (C.24) and (C.28), we can compute the derivative of this w.r.t. $\boldsymbol{\Lambda}_k$ and setting the result equal to zero, we find the MAP estimate for $\boldsymbol{\Lambda}_k$ to be

$$\boldsymbol{\Lambda}_k^{-1} = \frac{1}{\nu_0 + N_k - D}(\mathbf{W}_0^{-1} + \beta_0(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + N_k\mathbf{S}_k).$$

From this we see that $|\boldsymbol{\Lambda}_k^{-1}|$ 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.

10.29 Stardard 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

$$\begin{aligned} g(\lambda) &= \min_x \{\lambda x - f(x)\} \\ &= \min_x \{\lambda x - \ln(x)\}. \end{aligned}$$

We can minimize this w.r.t. x by setting the corresponding derivative to zero and solving for x:

$$\frac{dg}{dx} = \lambda - \frac{1}{x} = 0 \quad \Longrightarrow \quad x = \frac{1}{\lambda}.$$

Substituting this in (10.133), we see that

$$g(\lambda) = 1 - \ln\left(\frac{1}{\lambda}\right).$$

If we substitute this into (10.132), we get

$$f(x) = \min_\lambda \left\{\lambda x - 1 + \ln\left(\frac{1}{\lambda}\right)\right\}.$$

Again, we can minimize this w.r.t. λ by setting the corresponding derivative to zero and solving for λ:

$$\frac{df}{d\lambda} = x - \frac{1}{\lambda} = 0 \quad \Longrightarrow \quad \lambda = \frac{1}{x},$$

and substituting this into (10.132), we find that

$$f(x) = \frac{1}{x}x - 1 + \ln\left(\frac{1}{1/x}\right) = \ln(x).$$

10.32 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 directly 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.

10.37 Here we use the general expectation-propagation equations (10.204)–(10.207). The initial $q(\boldsymbol{\theta})$ takes the form

$$q_{\mathrm{init}}(\boldsymbol{\theta}) = \widetilde{f}_0(\boldsymbol{\theta})\prod_{i\neq 0}\widetilde{f}_i(\boldsymbol{\theta})$$

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 $q^{\mathrm{new}}(\boldsymbol{\theta})$ is determined by matching moments (sufficient statistics) against

$$q^{\backslash 0}(\boldsymbol{\theta})f_0(\boldsymbol{\theta}) = q_{\mathrm{init}}(\boldsymbol{\theta}).$$

Snce 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

$$Z_0 = \int q^{\backslash 0}(\boldsymbol{\theta}) f_0(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} = \int q^{\text{new}}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} = 1.$$

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

$$\begin{aligned} \text{var}\left[\widehat{f}\right] &= \mathbb{E}\left[\left(\widehat{f} - \mathbb{E}\left[\widehat{f}\right]\right)^2\right] \\ &= \mathbb{E}\left[\widehat{f}^2\right] - \mathbb{E}\left[f\right]^2. \end{aligned}$$

Now note

$$\begin{aligned} \mathbb{E}\left[f(z^{(k)}), f(z^{(m)})\right] &= \begin{cases} \text{var}[f] + \mathbb{E}[f^2] & \text{if } n = k, \\ \mathbb{E}[f^2] & \text{otherwise,} \end{cases} \\ &= \mathbb{E}[f^2] + \delta_{mk}\text{var}[f], \end{aligned}$$

where we again exploited the fact that the samples are independent.

Hence

$$\begin{aligned} \text{var}\left[\widehat{f}\right] &= \mathbb{E}\left[\frac{1}{L}\sum_{m=1}^{L} f(z^{(m)}) \frac{1}{L}\sum_{k=1}^{L} f(z^{(k)})\right] - \mathbb{E}[f]^2 \\ &= \frac{1}{L^2}\sum_{m=1}^{L}\sum_{k=1}^{L} \left\{\mathbb{E}[f^2] + \delta_{mk}\text{var}[f]\right\} - \mathbb{E}[f]^2 \\ &= \frac{1}{L}\text{var}[f] \\ &= \frac{1}{L}\mathbb{E}\left[(f - \mathbb{E}\left[f\right])^2\right]. \end{aligned}$$

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 trivially from the mechanism used to accept or reject the sample. The probability of a sample u drawn uniformly from the interval $[0, kq(\mathbf{z})]$ being less than or equal to a value $\widetilde{p}(\mathbf{z}) \leqslant kq(\mathbf{z})$ is simply

$$p(\text{acceptance}|\mathbf{z}) = \int_0^{\widetilde{p}(\mathbf{z})} \frac{1}{kq(\mathbf{z})}\,\mathrm{d}u = \frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})}.$$

Therefore, the probability density for drawing a sample, $\mathbf{z}$, is

$$q(\mathbf{z})p(\text{acceptance}|\mathbf{z}) = q(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})} = \frac{\widetilde{p}(\mathbf{z})}{k}. \tag{161}$$

Since $\widetilde{p}(\mathbf{z})$ is proportional to $p(\mathbf{x})$,

$$p(\mathbf{z}) = \frac{1}{Z_{\widetilde{p}}}\widetilde{p}(\mathbf{z}),$$

where

$$Z_{\widetilde{p}} = \int \widetilde{p}(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

As the l.h.s. of (161) is a probability density that integrates to 1, it follows that

$$\int \frac{\widetilde{p}(\mathbf{z})}{k}\,\mathrm{d}\mathbf{z} = 1$$

and so $k = Z_{\widetilde{p}}$, and

$$\frac{\widetilde{p}(\mathbf{z})}{k} = p(\mathbf{z}),$$

as required.

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 $T(\mathbf{z}, \mathbf{z}') = p^\star(z_k'|\{z_i\}_{i\neq k})$.

11.15 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

$$\frac{\partial H}{\partial z_i} = \frac{\partial E}{\partial r_i},$$

and from this, it is immediately clear that (11.55) and (11.59) are equivalent.

11.17 **NOTE**: In the first printing of PRML, there were 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 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 $M \leqslant D$.

12.4 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.

12.6 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.

12.8 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).

12.11 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).

12.15 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.

12.17 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

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n - \frac{1}{N}\sum_{n=1}^{N}\mathbf{W}\mathbf{z}_n = \overline{\mathbf{x}} - \mathbf{W}\overline{\mathbf{z}}.$$

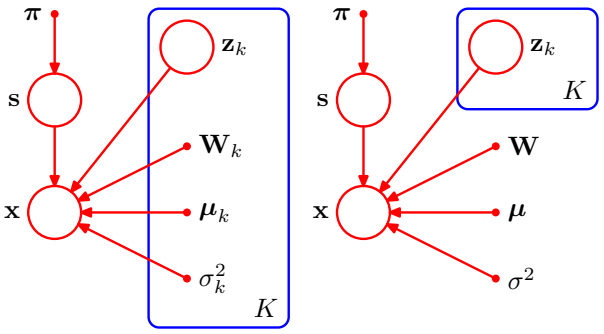


Figure 6 The left plot shows the graphical model corresponding to the general mixture of probabilistic PCA. The right plot shows the corresponding model were the parameter of all probabilist PCA models ($\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2) are shared across components. In both plots, $\mathbf{s}$ denotes the K-nomial latent variable that selects mixture components; it is governed by the parameter, $\boldsymbol{\pi}$.

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

$$J = \text{Tr}\left\{(\mathbf{X} - \mathbf{Z}\mathbf{W}^{\text{T}})(\mathbf{X} - \mathbf{Z}\mathbf{W}^{\text{T}})^{\text{T}}\right\}.$$

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).

12.19 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 $\mathbf{z}^{\text{T}}\mathbf{z} = \widetilde{\mathbf{z}}^{\text{T}}\widetilde{\mathbf{z}}$, where we have used $\mathbf{R}^{\text{T}}\mathbf{R} = \mathbf{I}$. 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 $\mathbf{W}\mathbf{W}^{\text{T}} = \widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^{\text{T}}$ and hence is also invariant to different choices of $\mathbf{R}$.

12.23 The solution is given in figure 6. 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.

12.25 Following the discussion of section 12.2, the log likelihood function for this model

can be written as

$$L(\boldsymbol{\mu}, \mathbf{W}, \boldsymbol{\Phi}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Phi}| - \frac{1}{2}\sum_{n=1}^{N}\left\{(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Phi})^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\right\},$$

where we have used (12.43).

If we consider the log likelihood function for the transformed data set we obtain

$$L_{\mathbf{A}}(\boldsymbol{\mu}, \mathbf{W}, \boldsymbol{\Phi}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Phi}| - \frac{1}{2}\sum_{n=1}^{N}\left\{(\mathbf{A}\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Phi})^{-1}(\mathbf{A}\mathbf{x}_n - \boldsymbol{\mu})\right\}.$$

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

$$L_{\mathbf{A}}(\boldsymbol{\mu}, \mathbf{W}, \boldsymbol{\Phi}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Phi}| - \frac{1}{2}\sum_{n=1}^{N}\mathrm{Tr}\left\{(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Phi})^{-1}\mathbf{A}\mathbf{S}\mathbf{A}^{\mathrm{T}}\right\}.$$

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 $\mathbf{\Phi}$ will be preserved by this re-scaling. In the case of probabilistic PCA the noise covariance $\mathbf{\Phi}$ 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 $\mathbf{\Phi}$ 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 7.

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

$$p(y) = q(f^{-1}(y)) \left| \frac{\mathrm{d}f^{-1}}{\mathrm{d}y} \right|. \tag{162}$$

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 8, as a part of Solution 1.4. From (162) we see directly that

$$|f'(x)| = \frac{q(x)}{p(f(x))}.$$

12.29 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

$$\bar{z}_i = \mathbb{E}[z_i] = \int z_i p(z_i)\, \mathrm{d}z_i.$$

NOTE: In the first printing of PRML, this exercise contained 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 $y_2 = y_2^2$).

Then we have

$$p(y_2|y_1) = \delta(y_2 - y_1^2),$$

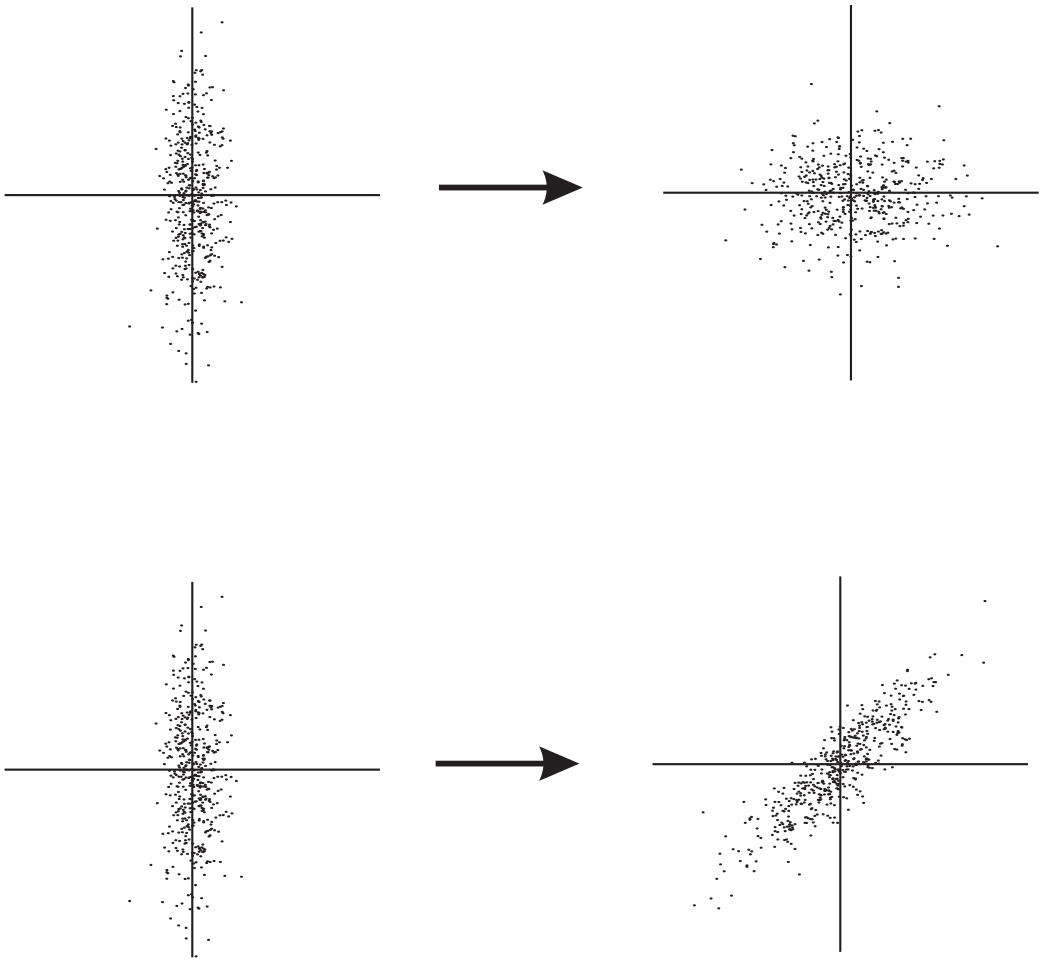


Figure 7 Factor analysis is covariant under a componentwise re-scaling of the data variables (top plots), while PCA and probabilistic PCA are covariant under rotations of the data space coordinates (lower plots).

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 and hence $\bar{y}_1$.

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}.

13.4 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.

13.8 Only the final term of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}$ 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.

13.13 Using (8.64), we can rewrite (13.50) as

$$\alpha(\mathbf{z}_n) = \sum_{\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}} F_n(\mathbf{z}_n, \{\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}\}), \tag{163}$$

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

$$F_n(\mathbf{z}_n, \{\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}\}) = h(\mathbf{z}_1) \prod_{i=2}^{n} f_i(\mathbf{z}_i, \mathbf{z}_{i-1}), \tag{164}$$

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 (164) as

$$F_n(\mathbf{z}_n, \{\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}\}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_2).$$

Applying the sum rule, summing over $\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}$ as on the r.h.s. of (163), we obtain (13.34).

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

$$\begin{aligned} h(\mathbf{z}_1) &= p(\mathbf{z}_1|\mathbf{u}_1)p(\mathbf{x}_1|\mathbf{z}_1, \mathbf{u}_1) \\ f_n(\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{u}_n)p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{u}_n). \end{aligned}$$

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.

13.22 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).

13.24 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 $\mathbf{c}$.

13.27 **NOTE**: In the first 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 $\mathbf{x}_n = \mathbf{C}\mathbf{z}_n$.

13.32 We can write the expected complete log-likelihood, given by the equation after (13.109), as a function of $\boldsymbol{\mu}_0$ and $\mathbf{V}_0$, as follows:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = &-\frac{1}{2}\ln|\mathbf{V}_0| \\ &-\frac{1}{2}\mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}^{\text{old}}}\left[\mathbf{z}_1^{\mathrm{T}}\mathbf{V}_0^{-1}\mathbf{z}_1 - \mathbf{z}_1^{\mathrm{T}}\mathbf{V}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^{\mathrm{T}}\mathbf{V}_0^{-1}\mathbf{z}_1 + \boldsymbol{\mu}_0^{\mathrm{T}}\mathbf{V}_0^{-1}\boldsymbol{\mu}_0\right] \qquad (165) \\ = &\frac{1}{2}\left(\ln|\mathbf{V}_0^{-1}| - \mathrm{Tr}\left[\mathbf{V}_0^{-1}\mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}^{\text{old}}}\left[\mathbf{z}_1\mathbf{z}_1^{\mathrm{T}} - \mathbf{z}_1\boldsymbol{\mu}_0^{\mathrm{T}} - \boldsymbol{\mu}_0\mathbf{z}_1^{\mathrm{T}} + \boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathrm{T}}\right]\right]\right), \qquad (166) \end{aligned}$$

where we have used (C.13) and omitted terms independent of $\boldsymbol{\mu}_0$ and $\mathbf{V}_0$.

From (165), we can calculate the derivative w.r.t. $\boldsymbol{\mu}_0$ using (C.19), to get

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_0} = 2\mathbf{V}_0^{-1}\boldsymbol{\mu}_0 - 2\mathbf{V}_0^{-1}\mathbb{E}[\mathbf{z}_1].$$

Setting this to zero and rearranging, we immediately obtain (13.110).

Using (166), (C.24) and (C.28), we can evaluate the derivatives w.r.t. $\mathbf{V}_0^{-1}$,

$$\frac{\partial Q}{\partial \mathbf{V}_0^{-1}} = \frac{1}{2}\left(\mathbf{V}_0 - \mathbb{E}[\mathbf{z}_1\mathbf{z}_1^{\mathrm{T}}] - \mathbb{E}[\mathbf{z}_1]\boldsymbol{\mu}_0^{\mathrm{T}} - \boldsymbol{\mu}_0\mathbb{E}[\mathbf{z}_1^{\mathrm{T}}] + \boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathrm{T}}\right).$$

Setting this to zero, rearrangning and making use of (13.110), we get (13.111).

Chapter 14 Combining Models

14.1 The required predictive distribution is given by

$$\begin{aligned} p(\mathbf{t}|\mathbf{x}, \mathbf{X}, \mathbf{T}) = &\\ &\sum_h p(h) \sum_{\mathbf{z}_h} p(\mathbf{z}_h) \int p(\mathbf{t}|\mathbf{x}, \boldsymbol{\theta}_h, \mathbf{z}_h, h) p(\boldsymbol{\theta}_h|\mathbf{X}, \mathbf{T}, h)\, \mathrm{d}\boldsymbol{\theta}_h, \qquad (167) \end{aligned}$$

where

$$\begin{aligned} p(\boldsymbol{\theta}_h|\mathbf{X},\mathbf{T},h) &= \frac{p(\mathbf{T}|\mathbf{X},\boldsymbol{\theta}_h,h)p(\boldsymbol{\theta}_h|h)}{p(\mathbf{T}|\mathbf{X},h)} \\ &\propto p(\boldsymbol{\theta}|h)\prod_{n=1}^{N} p(\mathbf{t}_n|\mathbf{x}_n,\boldsymbol{\theta},h) \\ &= p(\boldsymbol{\theta}|h)\prod_{n=1}^{N}\left(\sum_{\mathbf{z}_{nh}} p(\mathbf{t}_n,\mathbf{z}_{nh}|\mathbf{x}_n,\boldsymbol{\theta},h)\right) \end{aligned} \tag{168}$$

The integrals and summations in (167) 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 (168), 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.

14.3 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).

14.5 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 $y_{\mathrm{COM}}(\mathbf{x})$ as a result. The maximum value of $y_{\mathrm{COM}}(\mathbf{x})$ 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 $\boldsymbol{\alpha}$,

$$y_{\min}(\mathbf{x}) < y_{\mathrm{COM}}(\mathbf{x}) < y_{\max}(\mathbf{x}),$$

since $y_{\mathrm{COM}}(\mathbf{x})$ 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.

14.9 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.

14.13 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

$$p(t, \mathbf{z}) = p(t|\mathbf{z})p(\mathbf{z}) = \prod_{k=1}^{K} \left(\mathcal{N}\left(t \mid \mathbf{w}_k^{\mathrm{T}} \boldsymbol{\phi}, \beta^{-1} \right) \pi_k \right)^{z_k}.$$

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).

14.15 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.

14.17 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

$$p(t|\mathbf{x}) = \sum_{l=1}^{L} \eta_l \phi_l(t|\mathbf{x}),$$

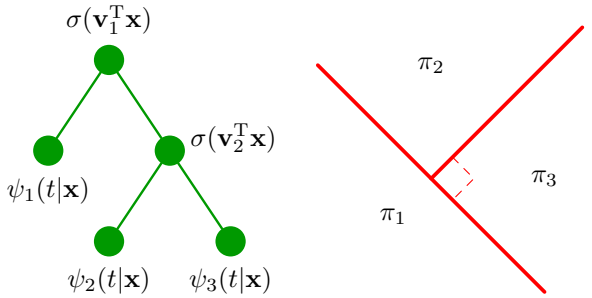


Figure 8 Left: an illustration of a hierarchical mixture model, where the input dependent mixing coefficients are determined by linear logistic models associated with interior nodes; the leaf nodes correspond to local (conditional) density models. Right: a possible division of the input space into regions where different mixing coefficients dominate, under the model illustrated left.

where $L = KM$, $l = (k-1)M + m$, $\eta_l = \pi_k \lambda_{mk}$ 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 8. 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

$$\begin{aligned}
\pi_1(\mathbf{x}) &= \sigma(\mathbf{v}_1^{\mathrm{T}}\mathbf{x}) \\
\pi_2(\mathbf{x}) &= (1 - \sigma(\mathbf{v}_1^{\mathrm{T}}\mathbf{x}))\sigma(\mathbf{v}_2^{\mathrm{T}}\mathbf{x}) \\
\pi_3(\mathbf{x}) &= (1 - \sigma(\mathbf{v}_1^{\mathrm{T}}\mathbf{x}))(1 - \sigma(\mathbf{v}_2^{\mathrm{T}}\mathbf{x})),
\end{aligned}$$

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 8, 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 8. 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 8.