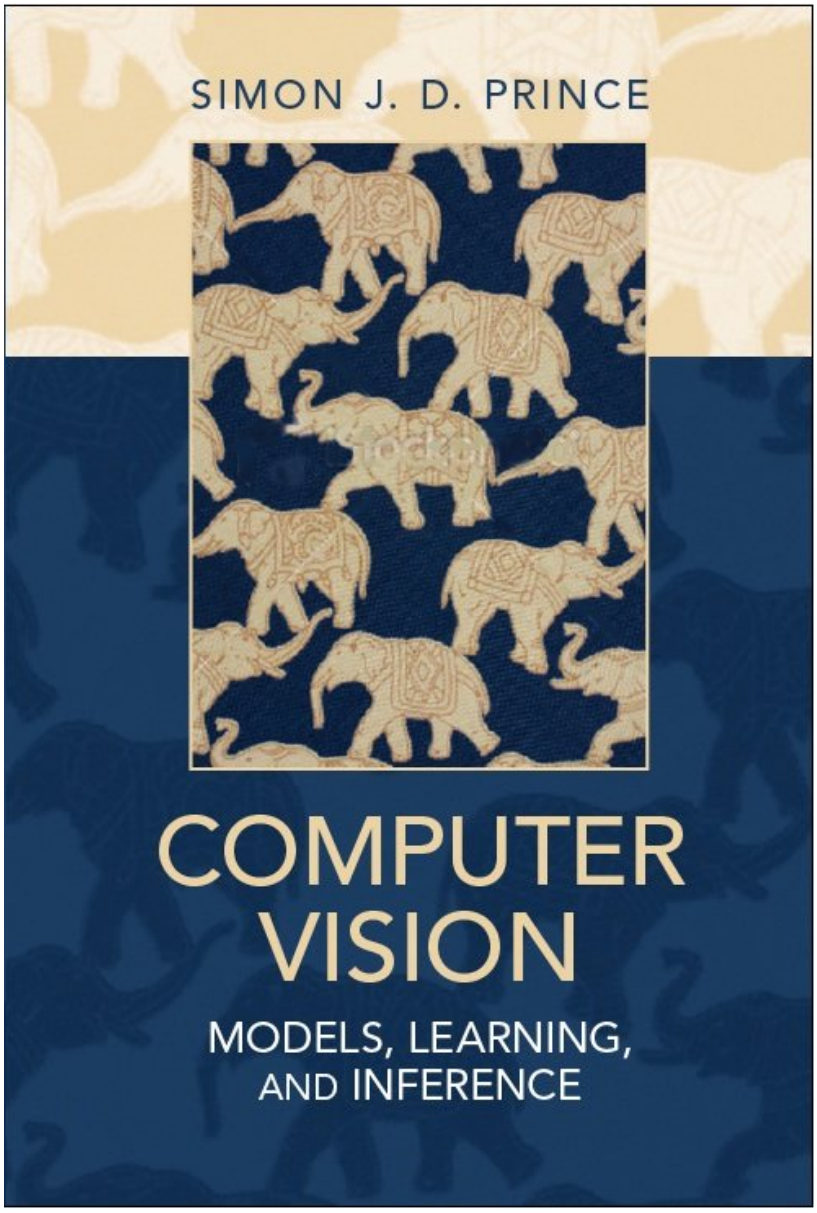

Answer booklet for students

July 7, 2012

Answer booklet

This document accompanies the book "Computer vision: models, learning, and inference" by Simon J.D. Prince, (http://www.computervisionmodels.com). This document contains answers to a selected subset of the problems at the end of each of the chapters of the main book. The remaining answers are available only to instructors via Cambridge University Press.

The included answers mostly involve derivations which would have detracted from the main text. I've also included the answers to any problems which contained an error in the original book, and I've corrected all of these errors in the current document. If you are systematically working through the problems in the book, it is hence better to work from this booklet which I will try to keep as up to date as possible.

This document has not yet been checked very carefully. I really need your help in this regard and I'd be very grateful if you would mail me at s.prince@cs.ucl.ac.uk if you cannot understand the text or if you think that you find a mistake. Suggestions for extra problems will also be gratefully received!

Simon Prince
July 7, 2012

Chapter 2

Introduction to probability

Problem 2.1 Give a real-world example of a joint distribution $Pr(x, y)$ where x is discrete and y is continuous.

Problem 2.2 What remains if I marginalize a joint distribution $Pr(v, w, x, y, z)$ over five variables with respect to variables w and y? What remains if I marginalize the resulting distribution with respect to v?

Problem 2.3 Show that the following relation is true:

$$Pr(w, x, y, z) = Pr(x, y)Pr(z|w, x, y)Pr(w|x, y).$$

Problem 2.4 In my pocket there are two coins. Coin 1 is unbiased, so the likelihood $Pr(h = 1|c = 1)$ of getting heads is 0.5 and the likelihood $Pr(h = 0|c = 1)$ of getting tails is also 0.5. Coin 2 is biased, so the likelihood $Pr(h = 1|c = 2)$ of getting heads is 0.8 and the likelihood $Pr(h = 0|c = 2)$ of getting tails is 0.2. I reach into my pocket and draw one of the coins at random. There is an equal prior probability I might have picked either coin. I flip the coin and observe a head. Use Bayes' rule to compute the posterior probability that I chose coin 2.

Problem 2.5 If variables x and y are independent and variables x and z are independent, does it follow that variables y and z are independent?

Answer

No, it does not follow. Consider any general distribution $Pr(y, z)$ where y and z are *NOT* independent. Now consider the marginal distributions $Pr(y)$ and $Pr(z)$. It is perfectly possible to have a third distribution $Pr(x)$ which does not provide any information about y or z and hence is independent of each and $Pr(x, y) = Pr(x)Pr(y)$ and $Pr(x, z) = Pr(x)Pr(z)$.

If you are unsure about this, then construct a counter example where x, y and z are all discrete variables with two entries. Construct a non-independent 2×2 distribution between y and z, marginalize it with respect to each and then construct 2×2 independent joint distributions between x and z and x and y.

Problem 2.6 Use equation 2.3 to show that when x and y are independent, the marginal distribution $Pr(x)$ is the same as the conditional distribution $Pr(x|y=y^*)$ for any y^*.

Problem 2.7 The joint probability $Pr(w,x,y,z)$ over four variables factorizes as

$$Pr(w,x,y,z) = Pr(w)Pr(z|y)Pr(y|x,w)Pr(x).$$

Demonstrate that x is independent of w by showing that $Pr(x,w) = Pr(x)Pr(w)$.

Problem 2.8 Consider a biased die where the probabilities of rolling sides $\{1,2,3,4,5,6\}$ are $\{1/12, 1/12, 1/12, 1/12, 1/6, 1/2\}$, respectively. What is the expected value of the die? If I roll the die twice, what is the expected value of the sum of the two rolls?

Problem 2.9 Prove the four relations for manipulating expectations.

$$\begin{aligned}
\mathrm{E}[\kappa] &= \kappa, \\
\mathrm{E}[\kappa f[x]] &= \kappa \mathrm{E}[f[x]], \\
\mathrm{E}[f[x]+g[x]] &= \mathrm{E}[f[x]] + \mathrm{E}[g[x]], \\
\mathrm{E}[f[x]g[y]] &= \mathrm{E}[f[x]]\mathrm{E}[g[y]], \qquad \text{if } x, y \text{ independent.}
\end{aligned}$$

Answer

Relation 1:

$$\begin{aligned}
E[\kappa] &= \int \kappa Pr(x)dx \\
&= \kappa \int Pr(x)dx \\
&= \kappa.
\end{aligned}$$

Relation 2:

$$\begin{aligned}
E[\kappa f[x]] &= \int \kappa f[x] Pr(x)dx \\
&= \kappa \int f[x] Pr(x)dx \\
&= \kappa E[f[x]].
\end{aligned}$$

Relation 3:

$$\begin{aligned}
E[f[x]+g[x]] &= \int (f[x]+g[x]) Pr(x)dx \\
&= \int \left(f[x]Pr(x) + g[x]Pr(x)\right) dx \\
&= \int f[x]Pr(x)dx + \int g[x]Pr(x)dx \\
&= E[f[x]] + E[g[x]].
\end{aligned}$$

Relation 4:

$$\begin{aligned} E[f[x]\cdot g[y]] &= \int\int f[x]\cdot g[y]Pr(x,y)dxdy \\ &= \int\int f[x]\cdot g[y]Pr(x)Pr(y)dxdy \\ &= \int f[x]Pr(x)dx\int g[y]Pr(y)dy \\ &= E[f[x]]E[g[y]], \end{aligned}$$

where we have used the definition of independence between the first two lines.

Problem 2.10 Use the relations from problem 2.9 to prove the following relationship between the second moment around zero and the second moment about the mean (variance):

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

Chapter 3

Common probability distributions

Problem 3.1 Consider a variable x which is Bernoulli distributed with parameter λ. Show that the mean $\mathrm{E}[x]$ is λ and the variance $\mathrm{E}[(x-\mathrm{E}[x])^2]$ is $\lambda(1-\lambda)$.

Problem 3.2 Calculate an expression for the mode (position of the peak) of the beta distribution with $\alpha, \beta > 1$ in terms of the parameters α and β.

Problem 3.3 The mean and variance of the beta distribution are given by the expressions

$$\begin{aligned} E[\lambda] = \mu &= \frac{\alpha}{\alpha+\beta} \\ E[(\lambda-\mu)^2] = \sigma^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

We may wish to choose the parameters α and β so that the distribution has a particular mean μ and variance σ^2. Derive suitable expressions for α and β in terms of μ and σ^2.

Problem 3.4 All of the distributions in this chapter are members of the *exponential family* and can be written in the form

$$Pr(x|\boldsymbol{\theta}) = a[x]\exp[\mathbf{b}[\theta]^T\mathbf{c}[x] - d[\theta]],$$

where $a[x]$ and $c[x]$ are functions of the data and $\mathbf{b}[\boldsymbol{\theta}]$ and $\mathbf{d}[\boldsymbol{\theta}]$ are functions of the parameters. Find the functions $a[x], \mathbf{b}[\theta], c[x]$ and $\mathbf{d}[\theta]$ that allow the Beta distribution to be represented in the generalized form of the exponential family.

Answer

The beta distribution is given by:

$$Pr(x) = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]}x^{\alpha-1}(1-x)^{\beta-1}.$$

We choose functions

$$\begin{aligned} a[x] &= 1 \\ \mathbf{b}[\boldsymbol{\theta}] &= \begin{bmatrix}\alpha-1\\ \beta-1\end{bmatrix} \\ \mathbf{c}[x] &= \begin{bmatrix}\log[x]\\ \log[1-x]\end{bmatrix} \\ d[\boldsymbol{\theta}] &= \log[\Gamma[\alpha]]+\log[\Gamma[\beta]]-\log[\Gamma[\alpha+\beta]], \end{aligned}$$

These functions now conform to the multivariate version of the exponential family.

$$Pr(x|\boldsymbol{\theta}) = a[x]\exp[\mathbf{b}[\theta]^T\mathbf{c}[x]-d[\theta]],$$

Problem 3.5 Use integration by parts to prove that if

$$\Gamma[z] = \int_0^\infty t^{z-1}e^{-t}dt,$$

then

$$\Gamma[z+1] = z\Gamma[z].$$

Problem 3.6 Consider a restricted family of univariate normal distributions where the variance is always 1, so that

$$Pr(x|\mu) = \frac{1}{\sqrt{2\pi}}\exp\left[-0.5(x-\mu)^2\right].$$

Show that a normal distribution over the parameter μ

$$Pr(\mu) = \text{Norm}_\mu[\mu_p, \sigma_p^2]$$

has a conjugate relationship to the restricted normal distribution.

Answer

Taking the product of the two distributions we get

$$\begin{aligned} Pr(x|\mu)\cdot Pr(\mu) &= \frac{1}{\sqrt{2\pi}}\exp\left[-0.5(x-\mu)^2\right]\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-0.5\frac{(\mu-\mu_p)^2}{\sigma_p^2}\right] \\ &= \frac{1}{2\pi\sigma}\exp\left[-0.5\left(x^2-2\mu x+\mu^2-\frac{\mu^2}{\sigma_p^2}-2\frac{2\mu\mu_p}{\sigma_p^2}+\frac{\mu_p^2}{\sigma_p^2}\right)\right] \\ &= \frac{1}{2\pi\sigma}\exp\left[-0.5\left(x^2+\frac{\mu_p^2}{\sigma_p^2}\right)\right]\exp\left[-0.5\left(\frac{\sigma_p^2+1}{\sigma_p^2}\mu^2-2\frac{\sigma_p^2x+\mu_p}{\sigma_p^2}\mu\right)\right] \\ &= \kappa_1\cdot\exp\left[-0.5\left(\frac{\sigma_p^2+1}{\sigma_p^2}\mu^2-2\frac{\sigma_p^2x+\mu_p}{\sigma_p^2}\mu\right)\right] \end{aligned}$$

The constant and the first exponential terms in the third line do not depend on μ so we denote them by the constant κ_1 for short. The second term is a quadratic function in μ.

We now complete the square of the second term by multiply and dividing by a suitable factor:

$$\begin{aligned}
Pr(x|\mu)\cdot Pr(\mu) &= \kappa_1 \cdot \exp\left[-0.5\left(\frac{\sigma_p^2+1}{\sigma_p^2}\mu^2 - 2\frac{\sigma_p^2 x+\mu_p}{\sigma_p^2}\mu\right)\right] \\
&= \kappa_1 \cdot \exp\left[-0.5\left(-\frac{\sigma_p^2}{\sigma_p^2+1}\left(\frac{\sigma_p^2 x+\mu_p}{\sigma_p^2}\right)^2\right)\right] \\
&\quad \cdot \exp\left[-0.5\left(\frac{\sigma_p^2+1}{\sigma_p^2}\mu^2 - 2\frac{\sigma_p^2 x+\mu_p}{\sigma_p^2}\mu + \frac{\sigma_p^2}{\sigma_p^2+1}\left(\frac{\sigma_p^2 x+\mu_p}{\sigma_p^2}\right)^2\right)\right] \\
&= \kappa_2 \cdot \exp\left[-0.5\left(\frac{\sigma_p^2+1}{\sigma_p^2}\mu^2 - 2\frac{\sigma_p^2 x+\mu_p}{\sigma_p^2}\mu + \frac{\sigma_p^2}{\sigma_p^2+1}\left(\frac{\sigma_p^2 x+\mu_p}{\sigma_p^2}\right)^2\right)\right] \\
&= \kappa_2 \cdot \exp\left[-0.5\frac{\left(\mu - \frac{\sigma_p^2 x+\mu_p}{\sigma_p^2+1}\right)^2}{\sigma_p^2/(\sigma_p^2+1)}\right]
\end{aligned}$$

where the extraneous factor that we multiplied by in the second line is incorporated into a second constant κ_2 between the second and third lines. The final result is a constant multiplied by a normal distribution and so we conclude that this distribution was conjugate as was originally claimed in the question.

Problem 3.7 For the normal distribution, find the functions $a[x], \mathbf{b}[\boldsymbol{\theta}], c[x]$ and $\mathbf{d}[\theta]$ that allow it to be represented in the generalized form of the exponential family (see problem 3.4).

Answer

The normal distribution has the form

$$\begin{aligned}
Pr(x) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-0.5\frac{(x-\mu)^2}{\sigma^2}\right] \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-0.5\left(\frac{x^2}{\sigma^2}+\frac{\mu^2}{\sigma^2}-2\frac{\mu x}{\sigma^2}\right)\right]
\end{aligned}$$

$$\begin{aligned}
a[x] &= 1 \\
\mathbf{b}[\boldsymbol{\theta}] &= \begin{bmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2} \end{bmatrix} \\
\mathbf{c}[x] &= \begin{bmatrix} x^2 \\ x \end{bmatrix} \\
d[\boldsymbol{\theta}] &= \frac{\mu^2}{2\sigma^2} + \log\left[\sqrt{2\pi\sigma^2}\right]
\end{aligned}$$

These functions now conform to the multivariate version of the exponential family.

$$Pr(x|\boldsymbol{\theta}) = a[x]\exp[\mathbf{b}[\theta]^T\mathbf{c}[x] - d[\theta]],$$

Problem 3.8 Calculate an expression for the mode (position of the peak in μ, σ^2 space) of the normal scaled inverse gamma distribution in terms of the parameters $\alpha, \beta, \gamma, \delta$.

Problem 3.9 Show that the more general form of the conjugate relation in which we multiply I Bernoulli distributions by the conjugate beta prior is given by

$$\prod_{i=1}^{I} \text{Bern}_{x_i}[\lambda] \cdot \text{Beta}_\lambda[\alpha, \beta] = \kappa \cdot \text{Beta}_\lambda[\tilde{\alpha}, \tilde{\beta}],$$

where

$$\begin{aligned}
\kappa &= \frac{\Gamma[\alpha+\beta]\Gamma[\alpha+\sum x_i]\Gamma[\beta+\sum(1-x_i)]}{\Gamma[\alpha+\beta+I]\Gamma[\alpha]\Gamma[\beta]} \\
\tilde{\alpha} &= \alpha + \sum x_i \\
\tilde{\beta} &= \beta + \sum(1-x_i).
\end{aligned}$$

Answer

The answer is achieved by multiplying out the terms to create an expression that is proportional to a new Beta distribution. We then multiply and divide by the constant associated with the new distribution.

$$\begin{aligned}
&\prod_{i=1}^{I} \text{Bern}_{x_i}[\lambda]\text{Beta}_\lambda[\alpha, \beta] \\
&= \prod_{i=1}^{I} \lambda^{x_i}(1-\lambda)^{1-x_i} \cdot \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]}\lambda^{\alpha-1}(1-\lambda)^{\beta-1} \\
&= \lambda^{\sum_i x_i}(1-\lambda)^{\sum_i(1-x_i)} \cdot \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]}\lambda^{\alpha-1}(1-\lambda)^{\beta-1} \\
&= \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]}\lambda^{\sum_i x_i+\alpha-1}(1-\lambda)^{I-\sum_i x_i+\beta-1} \\
&= \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]} \cdot \frac{\Gamma[\sum_i x_i+\alpha]\Gamma[I-\sum_i x_i+\beta]}{\Gamma[\sum x_i+\alpha+I-\sum x_i+\beta]}\text{Beta}_\lambda\left[\sum_i x_i+\alpha, I-\sum_i x_i+\beta\right] \\
&= \frac{\Gamma[\alpha+\beta]\Gamma[\sum_i x_i+\alpha]\Gamma[I-\sum_i x_i+\beta]}{\Gamma[\alpha+\beta+I]\Gamma[\alpha]\Gamma[\beta]}\text{Beta}_\lambda\left[\sum_i x_i+\alpha, I-\sum_i x_i+\beta\right]
\end{aligned}$$

as required.

Problem 3.10 Prove the conjugate relation

$$\prod_{i=1}^{I} \text{Cat}_{\mathbf{x}_i}[\lambda_{1\ldots K}] \cdot \text{Dir}_{\lambda_{1\ldots K}}[\alpha_{1\ldots K}] = \kappa \cdot \text{Dir}_{\lambda_{1\ldots K}}[\tilde{\alpha}_{1\ldots K}],$$

where

$$\begin{aligned} \tilde{\kappa} &= \frac{\Gamma[\sum_{j=1}^{K}\alpha_j]}{\Gamma[I+\sum_{j=1}^{K}\alpha_j]} \cdot \frac{\prod_{j=1}^{K}\Gamma[\alpha_j+N_j]}{\prod_{j=1}^{K}\Gamma[\alpha_j]} \\ \tilde{\alpha}_{1\ldots K} &= [\alpha_1+N_1, \alpha_2+N_2, \ldots, \alpha_K+N_K]. \end{aligned}$$

and N_k is the total number of times that the variable took the value k.

Answer

$$\begin{aligned} &\prod_{i=1}^{I} \text{Cat}_{\mathbf{x}_i}[\lambda_{1\ldots K}]\text{Dir}_{\lambda_{1\ldots K}}[\alpha_{1\ldots K}] \\ &= \left(\prod_{i=1}^{I}\prod_{j=1}^{K}\lambda_j^{x_{ij}}\right)\frac{\Gamma[\sum_{j=1}^{K}\alpha_j]}{\prod_{j=1}^{K}\Gamma[\alpha_j]}\prod_{j=1}^{K}\lambda_j^{\alpha_j-1} \\ &= \frac{\Gamma[\sum_{j=1}^{K}\alpha_j]}{\prod_{j=1}^{K}\Gamma[\alpha_j]}\prod_{j=1}^{K}\lambda_j^{\alpha_j-1+N_j} \\ &= \frac{\Gamma[\sum_{j=1}^{K}\alpha_j]}{\prod_{j=1}^{K}\Gamma[\alpha_j]}\cdot\left(\frac{\prod_{j=1}^{K}\Gamma[\alpha_j+N_j]}{\Gamma[I+\sum_{j=1}^{K}\alpha_j]}\right)\left(\frac{\Gamma[I+\sum_{j=1}^{K}\alpha_j]}{\prod_{j=1}^{K}\Gamma[\alpha_j+N_j]}\right)\cdot\prod_{j=1}^{K}\lambda_j^{\alpha_j+N_j-1} \\ &= \frac{\Gamma[\sum_{j=1}^{K}\alpha_j]}{\prod_{j=1}^{K}\Gamma[\alpha_j]}\cdot\left(\frac{\prod_{j=1}^{K}\Gamma[\alpha_j+N_j]}{\Gamma[I+\sum_{j=1}^{K}\alpha_j]}\right)\text{Dir}_{\lambda_{1\ldots K}}[\alpha_1+N_1,\alpha_2+N_2,\ldots,\alpha_k+N_K] \\ &= \frac{\Gamma[\sum_{j=1}^{K}\alpha_j]}{\Gamma[I+\sum_{j=1}^{K}\alpha_j]}\cdot\frac{\prod_{j=1}^{K}\Gamma[\alpha_j+N_j]}{\prod_{j=1}^{K}\Gamma[\alpha_j]}\text{Dir}_{\lambda_{1\ldots K}}[\alpha_1+N_1,\alpha_2+N_2,\ldots,\alpha_k+N_K] \end{aligned}$$

where $N_j = \sum_i x_{ij}$ and $\sum_j N_j = I$. The new parameters and constant can easily be read off from this expression.

Problem 3.11 Show that the conjugate relation between the normal and normal inverse gamma is given by

$$\prod_{i=1}^{I} \text{Norm}_{x_i}[\mu,\sigma^2] \cdot \text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta] = \kappa \cdot \text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}],$$

where

$$\begin{aligned}
\kappa &= \frac{1}{(2\pi)^{I/2}}\frac{\sqrt{\gamma}\beta^{\alpha}}{\sqrt{\tilde{\gamma}}\tilde{\beta}^{\tilde{\alpha}}}\frac{\Gamma[\tilde{\alpha}]}{\Gamma[\alpha]} \\
\tilde{\alpha} &= \alpha + I/2 \\
\tilde{\beta} &= \frac{\sum_i x_i^2}{2} + \beta + \frac{\gamma\delta^2}{2} - \frac{(\gamma\delta+\sum_i x_i)^2}{2(\gamma+I)} \\
\tilde{\gamma} &= \gamma + I \\
\tilde{\delta} &= \frac{(\gamma\delta+\sum_i x_i)}{\gamma+I}.
\end{aligned}$$

Answer

$$\begin{aligned}
&\prod_{i=1}^{I}\text{Norm}_{x_i}[\mu,\sigma^2]\text{NormInvGam}_x[\alpha,\beta,\gamma,\delta] = \\
&= \frac{1}{(2\pi)^{I/2}\sigma^I}\frac{\sqrt{\gamma}\beta^{\alpha}}{\sigma\sqrt{2\pi}\Gamma[\alpha]}\left(\frac{1}{\sigma^2}\right)^{(\alpha+1)} \\
&\quad \exp\left[-\frac{1}{2\sigma^2}\left(\sum_i(x_i-\mu)^2 + 2\beta + \gamma(\delta-\mu)^2\right)\right] \\
&= \frac{1}{(2\pi)^{I/2}\sigma^I}\frac{\sqrt{\gamma}\beta^{\alpha}}{\sigma\sqrt{2\pi}\Gamma[\alpha]}\left(\frac{1}{\sigma^2}\right)^{(\alpha+1)} \\
&\quad \exp\left[-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 + I\mu^2 - 2\mu\sum_i x_i + 2\beta + \gamma\delta^2 + \gamma\mu^2 - 2\mu\gamma\delta\right)\right] \\
&= \frac{1}{(2\pi)^{I/2}\sigma^I}\frac{\sqrt{\gamma}\beta^{\alpha}}{\sigma\sqrt{2\pi}\Gamma[\alpha]}\left(\frac{1}{\sigma^2}\right)^{(\alpha+1)} \\
&\quad \exp\left[-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 + 2\beta + \gamma\delta^2 + (\gamma+I)\left(\mu - \frac{(\gamma\delta+\sum_i x_i)}{\gamma+I}\right)^2 - \frac{(\gamma\delta+\sum_i x_i)^2}{\gamma+I}\right)\right] \\
&= \frac{1}{(2\pi)^{I/2}\sigma^I}\frac{\sqrt{\gamma}\beta^{\alpha}}{\sigma\sqrt{2\pi}\Gamma[\alpha]}\left(\frac{1}{\sigma^2}\right)^{(\alpha+1)} \\
&\quad \exp\left[-\frac{1}{2\sigma^2}\left(2\left(\frac{\sum_i x_i^2}{2} + \beta + \frac{\gamma\delta^2}{2} - \frac{(\gamma\delta+\sum_i x_i)^2}{2(\gamma+I)}\right) + (\gamma+I)\left(\mu - \frac{(\gamma\delta+\sum_i x_i)}{\gamma+I}\right)^2\right)\right],
\end{aligned}$$

from which the parameters and the constant can easily be read off.

Problem 3.12 Show that the conjugate relationship between the multivariate normal and the normal inverse Wishart is given by

$$\prod_{i=1}^{I}\text{Norm}_{\mathbf{x}_i}\left[\boldsymbol{\mu},\boldsymbol{\Sigma}\right]\cdot\text{NorIWis}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left[\alpha,\boldsymbol{\Psi},\gamma,\boldsymbol{\delta}\right] = \kappa\cdot\text{NorIWis}\left[\tilde{\alpha},\tilde{\boldsymbol{\Psi}},\tilde{\gamma},\tilde{\boldsymbol{\delta}}\right],$$

where

$$
\begin{aligned}
\kappa &= \frac{1}{\pi^{ID/2}} \frac{\boldsymbol{\Psi}^{\alpha/2}}{\tilde{\boldsymbol{\Psi}}^{\tilde{\alpha}/2}} \frac{\Gamma_D[\tilde{\alpha}/2]}{\Gamma_D[\alpha/2]} \frac{\gamma^{D/2}}{\tilde{\gamma}^{D/2}} \\
\tilde{\alpha} &= \alpha + I \\
\tilde{\boldsymbol{\Psi}} &= \boldsymbol{\Psi} + \gamma\boldsymbol{\delta}\boldsymbol{\delta}^T + \sum_{i=1}^{I} \mathbf{x}_i \mathbf{x}_i^T - \frac{1}{(\gamma+I)} \left(\gamma\boldsymbol{\delta} + \sum_{i=1}^{I} \mathbf{x}_i\right) \left(\gamma\boldsymbol{\delta} + \sum_{i=1}^{I} \mathbf{x}_i\right)^T \\
\tilde{\gamma} &= \gamma + I \\
\tilde{\boldsymbol{\delta}} &= \frac{\gamma\boldsymbol{\delta} + \sum_{i=1}^{I} \mathbf{x}_i}{\gamma + I}.
\end{aligned}
$$

You may need to use the relation $\text{Tr}\left[\mathbf{z}\mathbf{z}^T\mathbf{A}^{-1}\right] = \mathbf{z}^T\mathbf{A}^{-1}\mathbf{z}$.

Answer

TO DO

Chapter 4

Fitting probability models

Problem 4.1 Show that the maximum likelihood solution for the variance σ^2 of the normal distribution is given by

$$\sigma^2 = \sum_{i=1}^{I} \frac{(x_i - \hat{\mu})^2}{I}.$$

Problem 4.2 Show that the MAP solution for the mean μ and variance σ^2 of the normal distribution are given by

$$\hat{\mu} = \frac{\sum_{i=1}^{I} x_i + \gamma\delta}{I+\gamma} \qquad \text{and} \qquad \hat{\sigma^2} = \frac{\sum_{i=1}^{I}(x_i - \hat{\mu})^2 + 2\beta + \gamma(\delta - \hat{\mu})^2}{I+3+2\alpha},$$

when we use the conjugate normal-scaled inverse gamma prior

$$Pr(\mu, \sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma[\alpha]} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta-\mu)^2}{2\sigma^2}\right].$$

Problem 4.3 Taking equation 4.29 as a starting point, show that the maximum likelihood parameters for the categorical distribution are given by

$$\hat{\lambda}_k = \frac{N_k}{\sum_{m=1}^{6} N_m},$$

where N_k is the number of times that category K was observed in the training data.

Problem 4.4 Show that the MAP estimate for the parameters $\{\lambda\}_{k=1}^{K}$ of the categorical distribution is given by

$$\hat{\lambda}_k = \frac{N_k + \alpha_k - 1}{\sum_{m=1}^{6}(N_m + \alpha_m - 1)},$$

under the assumption of a Dirichlet prior with hyperparameters $\{\alpha_k\}_{k=1}^{K}$. The terms N_k again indicate the number of times that category k was observed in the training data.

Problem 4.5 The denominator of Bayes' rule

$$Pr(x_{1\ldots I}) = \int \prod_{i=1}^{I} Pr(x_i|\theta) Pr(\theta)\, d\theta$$

is known as the *evidence*. It is a measure of how well the distribution fits *regardless* of the particular values of the parameters. Find an expression for the evidence term for (i) the normal distribution and (ii) the categorical distribution assuming conjugate priors in each case.

Problem 4.6 The evidence term can be used to compare models. Consider two sets of data $\mathcal{S}_1 = \{0.1, -0.5, 0.2, 0.7\}$ and $\mathcal{S}_2 = \{1.1, 2.0, 1.4, 2.3\}$. Let us pose the question of whether these two data sets came from the same normal distribution or from two different normal distributions.

Let model M_1 denote the case where all of the data comes from the one normal distribution. The evidence for this model is

$$Pr(\mathcal{S}_1 \cup \mathcal{S}_2|M_1) = \int \prod_{i \in \mathcal{S}_1 \cup \mathcal{S}_2} Pr(x_i|\boldsymbol{\theta}) Pr(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

where $\boldsymbol{\theta} = \{\mu, \sigma^2\}$ contains the parameters of this normal distribution. Similarly, we will let M_2 denote the case where the two sets of data belong to different normal distributions

$$Pr(\mathcal{S}_1 \cup \mathcal{S}_2|M_2) = \int \prod_{i \in \mathcal{S}_1} Pr(x_i|\boldsymbol{\theta}_1) Pr(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1 \int \prod_{i \in \mathcal{S}_2} Pr(x_i|\boldsymbol{\theta}_2) Pr(\boldsymbol{\theta}_2)\, d\boldsymbol{\theta}_2,$$

where $\boldsymbol{\theta}_1 = \{\mu_1, \sigma_1^2\}$ and $\boldsymbol{\theta}_2 = \{\mu_2, \sigma_2^2\}$.

Now it is possible to compare the probability of the data under each of these two models using Bayes' rule

$$Pr(M_1|\mathcal{S}_1 \cup \mathcal{S}_2) = \frac{Pr(\mathcal{S}_1 \cup \mathcal{S}_2|M_1) Pr(M_1)}{\sum_{n=1}^{2} Pr(\mathcal{S}_1 \cup \mathcal{S}_2|M_n) Pr(M_n)}$$

Use this expression to compute the posterior probability that the two datasets came from the same underlying normal distribution. You may assume normal-scaled inverse gamma priors over $\boldsymbol{\theta}$, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ with parameters $\alpha = 1, \beta = 1, \gamma = 1, \delta = 0$.

Answer

From Problem 4.5 part one, we have that

$$Pr(x_{1\ldots I}) = \kappa = \frac{1}{(2\pi)^{I/2}} \frac{\sqrt{\gamma}\beta^{\alpha}}{\sqrt{\tilde{\gamma}}\tilde{\beta}^{\tilde{\alpha}}}$$

where

$$\begin{aligned}
\tilde{\alpha} &= \alpha + I/2 \\
\tilde{\beta} &= \frac{\sum_i x_i^2}{2} + \beta + \frac{\gamma\delta^2}{2} - \frac{(\gamma\delta + \sum_i x_i)^2}{2(\gamma + I)} \\
\tilde{\gamma} &= \gamma + I
\end{aligned}$$

Now we simple compute this expression for the data $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_1 \cup \mathcal{S}_2$ and substitute into Bayes' rule.

The resulting probabilities are

$$\begin{aligned} Pr(\mathcal{S}_1) &= 4.4 \times 10^{-3} \\ Pr(\mathcal{S}_2) &= 6.4 \times 10^{-4} \\ Pr(\mathcal{S}_1 \cup \mathcal{S}_1) &= 9.7 * 10^{-8} \end{aligned}$$

and the posterior probability if the two sets of data coming from the same normal is

$$Pr(M_1|\mathcal{S}_1 \cup \mathcal{S}_2) = 0.0344,$$

Problem 4.7 In the Bernoulli distribution, the likelihood $Pr(x_{1\ldots I}|\lambda)$ of the parameter λ under the data $\{x_i\}_{i=1}^I$ where $x_i \in \{0, 1\}$ is

$$Pr(x_{1\ldots I}|\lambda) = \prod_{i=1}^{I} \lambda^{x_i}(1-\lambda)^{1-x_i}.$$

Find an expression for the maximum likelihood estimate of the parameter λ.

Problem 4.8 Find an expression for the MAP estimate of the Bernoulli parameter λ (see problem 4.7) assuming a beta distributed prior

$$Pr(\lambda) = \text{Beta}_\lambda[\alpha, \beta].$$

Problem 4.9 Now consider the Bayesian approach to fitting Bernoulli data, using a beta distributed prior. Find expressions for (i) the posterior probability distribution over the Bernoulli parameters given observed data $\{x_i\}_{i=1}^I$ and (ii) the predictive distribution for new data $\mathbf{x}^*$.

Problem 4.10 Staying with the Bernoulli distribution, consider observing data $0, 0, 0, 0$ from four trials. Assuming a uniform beta prior ($\alpha = 1, \beta = 1$), compute the predictive distribution using the (i) maximum likelihood, (ii) maximum a posteriori and (iii) Bayesian approaches. Comment on the results.

Chapter 5

The normal distribution

Problem 5.1 Consider a multivariate normal distribution in variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Show that if we make the linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ then the transformed variable $\mathbf{y}$ is distributed as:

$$Pr(\mathbf{y}) = \text{Norm}_{\mathbf{y}}\left[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\right].$$

Answer

The multivariate normal distribution is defined as

$$Pr(\mathbf{x}) = \frac{1}{(2\pi)^D|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-0.5\,(\mathbf{x}-\boldsymbol{\mu})\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\right]$$

To transform the distribution as $\mathbf{y} = \mathbf{A}\mathbf{x}+\mathbf{b}$, we substitute in $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})$ and divide by the determinant of the Jacobian of the transformation, which for this case is just $|\mathbf{A}|$ giving:

$$Pr(\mathbf{x}) = \frac{1}{(2\pi)^D|\boldsymbol{\Sigma}|^{1/2}|\mathbf{A}|} \exp\left[-0.5\,\left(\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu}\right)\,\boldsymbol{\Sigma}^{-1}\,\left(\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu}\right)\right]$$

We'll first work with the constant κ from outside the exponential to show that it is the constant for a distribution with variance $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$. We have

$$\begin{aligned}
\kappa &= \frac{1}{(2\pi)^D|\boldsymbol{\Sigma}|^{1/2}|\mathbf{A}|} \\
&= \frac{1}{(2\pi)^D|\boldsymbol{\Sigma}|^{1/2}|\mathbf{A}|^{1/2}|\mathbf{A}|^{1/2}} \\
&= \frac{1}{(2\pi)^D|\boldsymbol{\Sigma}|^{1/2}|\mathbf{A}|^{1/2}|\mathbf{A}^T|^{1/2}} \\
&= \frac{1}{(2\pi)^D|\mathbf{A}|^{1/2}|\boldsymbol{\Sigma}|^{1/2}|\mathbf{A}^T|^{1/2}} \\
&= \frac{1}{(2\pi)^D|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T|^{1/2}}
\end{aligned}$$

which is exactly the required constant.

Now we'll work with the quadratic in the exponential term to show that it corresponds to a normal distribution in $\mathbf{y}$ with variance $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ and mean $\mathbf{A}\boldsymbol{\mu}+\mathbf{b}$.

$$\begin{aligned}
&\left(\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu}\right)\\
&\quad=\mathbf{y}^T\mathbf{A}^{-T}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{y}-2(\mathbf{b}^T\mathbf{A}^{-T}+\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{y}+(\mathbf{A}^{-1}\mathbf{b}+\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{A}^{-1}\mathbf{b}+\boldsymbol{\mu})\\
&\quad=\mathbf{y}^T\mathbf{A}^{-T}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{y}-2(\mathbf{b}^T+\boldsymbol{\mu}^T\mathbf{A}^T)\mathbf{A}^{-T}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{y}\\
&\qquad\qquad+(\mathbf{A}^{-1}\mathbf{b}+\mathbf{A}\boldsymbol{\mu})^T\mathbf{A}^{-T}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}(\mathbf{b}+\mathbf{A}\boldsymbol{\mu})\\
&\quad=(\mathbf{y}-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))^T\left(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\right)^{-1}(\mathbf{y}-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))
\end{aligned}$$

This is clearly the quadratic term from a normal distribution in $\mathbf{y}$ with variance $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ and mean $\mathbf{A}\boldsymbol{\mu}+\mathbf{b}$ as required.

Problem 5.2 Show that we can convert a normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ to a new distribution with mean $\mathbf{0}$ and covariance $\mathbf{I}$ using the linear transformation $\mathbf{y}=\mathbf{A}\mathbf{x}+\mathbf{b}$ where

$$\begin{aligned}
\mathbf{A} &= \boldsymbol{\Sigma}^{-1/2}\\
\mathbf{b} &= -\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}.
\end{aligned}$$

This is known as the *whitening* transform.

Problem 5.3 Show that for multivariate normal distribution

$$Pr(\mathbf{x})=Pr\left(\begin{bmatrix}\mathbf{x}_1\\\mathbf{x}_2\end{bmatrix}\right)=\text{Norm}_{\mathbf{x}}\left[\begin{bmatrix}\boldsymbol{\mu}_1\\\boldsymbol{\mu}_2\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{11}&\boldsymbol{\Sigma}_{21}^T\\\boldsymbol{\Sigma}_{21}&\boldsymbol{\Sigma}_{22}\end{bmatrix}\right],$$

the marginal distribution in $\mathbf{x}_1$ is

$$Pr(\mathbf{x}_1) = \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11}\right].$$

Hint: apply the transformation $\mathbf{y}=[\mathbf{I},\mathbf{0}]\mathbf{x}$.

Answer

Using the transformation rule:

$$Pr(\mathbf{y})=\text{Norm}_{\mathbf{y}}\left[\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\right],$$

with $\mathbf{b}=\mathbf{0}$ and $\mathbf{A}=[\mathbf{I},\mathbf{0}]$,
we get

$$\begin{aligned}
Pr(\mathbf{x}_1) &= \text{Norm}_{\mathbf{x}_1}\left[[\mathbf{I},\mathbf{0}]\begin{bmatrix}\boldsymbol{\mu}_1\\\boldsymbol{\mu}_2\end{bmatrix}+\mathbf{0},[\mathbf{I},\mathbf{0}]\begin{bmatrix}\boldsymbol{\Sigma}_{11}&\boldsymbol{\Sigma}_{21}^T\\\boldsymbol{\Sigma}_{21}&\boldsymbol{\Sigma}_{22}\end{bmatrix}[\mathbf{I},\mathbf{0}]^T\right]\\
&= \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11}\right]
\end{aligned}$$

as required.

Problem 5.4 The Schur complement identity states that inverse of a matrix in terms of its sub-blocks is

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}^{-1} = \begin{bmatrix}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}\end{bmatrix}.$$

Show that this relation is true.

Answer

To show that this is true, we simply multiply the Schur matrix by the original matrix and show that the product equals the identity. In other words we compute

$$\begin{bmatrix}\mathbf{U} & \mathbf{V}\\ \mathbf{W} & \mathbf{X}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}\begin{bmatrix}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}\end{bmatrix}.$$

Now we need to show that $\mathbf{U}=\mathbf{I}$, $\mathbf{V}=\mathbf{0}$, $\mathbf{W}=\mathbf{0}$ and $\mathbf{X}=\mathbf{I}$, so that the left hand side as a whole is the identity matrix. Solving for $\mathbf{U}$ we have

$$\begin{aligned}\mathbf{U} &= \mathbf{A}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}+\mathbf{B}-\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\\ &= (\mathbf{A}+\mathbf{B}-\mathbf{D}^{-1}\mathbf{C})(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\\ &= \mathbf{I},\end{aligned}$$

as required. Solving for $\mathbf{V}$ we have

$$\begin{aligned}\mathbf{V} &= \mathbf{A}(-(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1})+\mathbf{B}\left(\mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}\right)\\ &= \mathbf{B}\mathbf{D}^{-1}-(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}\\ &= \mathbf{B}\mathbf{D}^{-1}-\mathbf{B}\mathbf{D}^{-1}\\ &= \mathbf{0},\end{aligned}$$

as required. Solving for $\mathbf{W}$ we get

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

as required. Solving for $\mathbf{X}$ we get

$$\begin{aligned}\mathbf{X} &= \mathbf{C}(-(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1})+\mathbf{D}(\mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1})\\ &= -\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}+\mathbf{I}+\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1})\\ &= \mathbf{I},\end{aligned}$$

as required.

Problem 5.5 Prove the conditional distribution property for the normal distribution: if

$$Pr(\mathbf{x}) = Pr\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\right) = \text{Norm}_{\mathbf{x}}\left[\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}^T\\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right],$$

then

$$Pr(\mathbf{x}_1|\mathbf{x}_2) = \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}\right].$$

Answer

We can write out the joint distribution as

$$Pr(\mathbf{x}) = \kappa_1 \exp\left[-0.5\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}\right)^T\begin{bmatrix}\mathbf{U} & \mathbf{V}\\ \mathbf{W} & \mathbf{X}\end{bmatrix}\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}\right)\right]$$

where κ_1 is the standard constant associated with the normal distribution and

$$\begin{aligned}\begin{bmatrix}\mathbf{U} & \mathbf{V}\\ \mathbf{W} & \mathbf{X}\end{bmatrix} &= \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}^T\\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22}\end{bmatrix}^{-1}\\ &= \begin{bmatrix}(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12})^{-1} & -(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12})^{-1} & \boldsymbol{\Sigma}_{22}^{-1}+\boldsymbol{\Sigma}_{22}^{-1}\mathbf{C}(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\end{bmatrix}.\end{aligned}$$

Now we exploit Bayes' rule $Pr(\mathbf{x}_1|\mathbf{x}_2) = Pr(\mathbf{x}_1,\mathbf{x}_2)/Pr(\mathbf{x}_2)$ and note that $Pr(\mathbf{x}_1|\mathbf{x}_2) \propto Pr(\mathbf{x}_1,\mathbf{x}_2)$. We hence take the approach of reformulating the joint distribution in terms of a normal distribution in $\mathbf{x}_1$ alone. Multiplying out the matrices in the exponent, we get

$$Pr(\mathbf{x}) = \kappa_2 \exp\left[-0.5\left((\mathbf{x}_1-\boldsymbol{\mu}_1)^T\mathbf{U}(\mathbf{x}_1-\boldsymbol{\mu}_1) + 2(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\mathbf{V}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right)\right]$$

where we have subsumed the terms that do not depend on $\mathbf{x}_1$ to create a new constant κ_2. Now we both add and subtract a term to complete the square,

$$\begin{aligned}Pr(\mathbf{x}) = \kappa_3 \exp\Big[-0.5\Big((\mathbf{x}_1-\boldsymbol{\mu}_1)^T\mathbf{U}(\mathbf{x}_1-\boldsymbol{\mu}_1)+2(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\mathbf{V}(\mathbf{x}_2-\boldsymbol{\mu}_2)\Big)\\ +(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\mathbf{V}^T\mathbf{U}^{-1}\mathbf{V}(\mathbf{x}_2-\boldsymbol{\mu}_2)\Big]\end{aligned}$$

where we have subsumed the term we must subtract in the exponential to balance the one we have added into the new constant κ_3. Re-factoring, we see that this is a normal distribution

$$Pr(\mathbf{x}) = \kappa_3 \exp\left[-0.5\left((\mathbf{x}_1-\boldsymbol{\mu}_1+\mathbf{U}^{-1}\mathbf{V}(\mathbf{x}_2-\boldsymbol{\mu}_2))^T\mathbf{U}(\mathbf{x}_1-\boldsymbol{\mu}_1+\mathbf{U}_{-1}\mathbf{V}(\mathbf{x}_2-\boldsymbol{\mu}_2))\right)\right]$$

The conditional distribution $Pr(\mathbf{x}_1|\mathbf{x}_2)$ is proportional to this expression, and the constant of proportionality must be $1/(\kappa_3|\mathbf{U}|^{1/2}2\pi^{D/2})$ so that it is a proper normal distribution. Hence the conditional distribution is given by

$$\begin{aligned} Pr(\mathbf{x}_1|\mathbf{x}_2) &= \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu}_1 - \mathbf{U}^{-1}\mathbf{V}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \mathbf{U}^{-1}\right] \\ &= \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}\right]. \end{aligned}$$

Problem 5.6 Use the conditional probability relation for the normal distribution to show that the conditional distribution $Pr(x_1|x_2 = k)$ is the same for all k when the covariance is diagonal and the variables are independent (see figure 5.5b).

Answer

The conditional probability relation is:

$$Pr(\mathbf{x}_1|\mathbf{x}_2) = \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}\right].$$

When the covariance is diagonal $\boldsymbol{\Sigma}_{12}$ which contains the off-diagonal elements relating $\mathbf{x}_1$ and $\mathbf{x}_2$ will be zero. Hence, in this case, we have

$$Pr(\mathbf{x}_1|\mathbf{x}_2) = \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}\right],$$

and the distribution is the same regardless of the value of $\mathbf{x}_2$.

Problem 5.7 Show that

$$\text{Norm}_{\mathbf{x}}[\mathbf{a}, \mathbf{A}]\text{Norm}_{\mathbf{x}}[\mathbf{b}, \mathbf{B}] \propto \text{Norm}_{\mathbf{x}}[(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b}), (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}].$$

Answer

We first write out the expression for the product of the normals:

$$\begin{aligned} \text{Norm}_{\mathbf{x}}\left[\mathbf{a}, \mathbf{A}\right]\text{Norm}_{\mathbf{x}}\left[\mathbf{b}, \mathbf{B}\right] &= \kappa_1 \exp\left[-0.5(\mathbf{x}-\mathbf{a})^T\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a}) - 0.5(\mathbf{x}-\mathbf{b})^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{b})\right] \\ &= \kappa_2 \exp\left[-0.5(\mathbf{x}^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{x} + 2(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\mathbf{x})\right], \end{aligned}$$

where we have multiplied out the terms in the exponential and subsumed the terms that are constant in $\mathbf{x}$ into a new constant κ_2.

Now we complete the square by multiplying and dividing by a new term, creating a new constant κ_3:

$$\begin{aligned} \text{Norm}_{\mathbf{x}}\left[\mathbf{a}, \mathbf{A}\right]\text{Norm}_{\mathbf{x}}\left[\mathbf{b}, \mathbf{B}\right] &= \kappa_2 \exp\left[-0.5(\mathbf{x}^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{x} + 2(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\mathbf{x})\right] \\ &= \kappa_3 \exp\left[-0.5(\mathbf{x}^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{x} + 2(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\mathbf{x}\right. \\ &\qquad \left. + (\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}))\right]. \end{aligned}$$

Finally, we re-factor the exponential term into the form of a standard normal distribution, and divide and multiply by the associated constant term for this new normal distribution, creating a new constant κ_4:

$$\begin{aligned}\text{Norm}_{\mathbf{x}}\left[\mathbf{a},\mathbf{A}\right]\text{Norm}_{\mathbf{x}}\left[\mathbf{b},\mathbf{B}\right] &= \kappa_3 \exp\Big[-0.5\left(\mathbf{x}-(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\right)^T\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)\\ &\qquad\qquad \left(\mathbf{x}-(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\right)\Big]\\ &= \kappa_4\text{Norm}_{\mathbf{x}}\left[(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}),(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\right].\end{aligned}$$

Problem 5.8 For the 1D case, show that when we take the product of the two normal distributions with means μ_1, μ_2 and variances σ_1^2, σ_2^2, the new mean lies between the original two means and the new variance is smaller than either of the original variances.

Answer

For the 1D case, the mean is given by

$$\mu = \frac{\sigma_1^{-2}\mu_1 + \sigma_2^{-2}}{\sigma_1^{-1} + \sigma_2^{-2}}$$

This has the form $\mu = a\mu_1 + b\mu_2$ where $a > 0$, $b > 0$ and $a + b = 1$ and so the new mean must be somewhere between the original two means.

The new variance is given by

$$\sigma^2 = \frac{1}{\sigma_1^{-1} + \sigma_2^{-2}}$$

To show that the new variance σ^2 is smaller than σ_1^2 we start with the uncontroversial proposition

$$\sigma_1^{-2} + \sigma_2^{-2} > \sigma_1^{-2},$$

where this must be true because all the terms are inverse variances and so are guaranteed to be positive. Now we take the reciprocal of both sides of the equation to yield

$$\frac{1}{\sigma_1^{-2} + \sigma_2^{-2}} < \sigma_1^2,$$

and since the left hand side is the equal to the new variance σ^2 we conclude that $\sigma^2 < \sigma_1^2$. A similar argument produces the symmetric result $\sigma^2 < \sigma_2^2$.

Problem 5.9 Show that the constant of proportionality κ in the product relation in problem 5.7 is also a normal distribution where

$$\kappa = \text{Norm}_{\mathbf{a}}[\mathbf{b}, \mathbf{A} + \mathbf{B}].$$

Answer

Using the answer from problem 5.7 for reference, we will compute each constant in turn.

$$\begin{aligned}
\kappa_1 &= \frac{1}{(2\pi)^D|\mathbf{A}|^{0.5}|\mathbf{B}|^{0.5}} \\
\kappa_2 &= \kappa_1 \cdot \exp\left[-0.5\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} - 0.5\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b}\right] \\
\kappa_3 &= \kappa_2 \cdot \exp\left[0.5(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\right] \\
\kappa_4 &= \kappa_3 \cdot 2\pi^{D/2}|(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}|^{0.5}
\end{aligned}$$

Working first with the constant, we have

$$\begin{aligned}
\kappa &= \frac{1}{(2\pi)^D|\mathbf{A}|^{0.5}|\mathbf{B}|^{0.5}} \cdot 2\pi^{D/2}|(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}|^{0.5}\exp[\bullet] \\
&= \frac{1}{2\pi^{D/2}|\mathbf{A}|^{0.5}||\mathbf{A}^{-1}+\mathbf{B}^{-1}|^{0.5}|\mathbf{B}|^{0.5}}\exp[\bullet] \\
&= \frac{1}{2\pi^{D/2}|\mathbf{A}+\mathbf{B}|^{0.5}}\exp[\bullet]
\end{aligned}$$

which is the appropriate constant for the normal distribution.
Now let's work on the exponential term. We have

$$\begin{aligned}
\kappa &\propto \exp\left[-0.5\left(\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}+\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b}-(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\right)\right] \\
&\propto \exp\Big[-0.5\Big(\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}+\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b}-\mathbf{a}^T\mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{A}^{-1}\mathbf{a} \\
&\qquad -\mathbf{b}^T\mathbf{B}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}\mathbf{b}-2\mathbf{a}^T\mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}\mathbf{b}\Big)\Big] \\
&\propto \exp\Big[-0.5\Big(\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}+\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b}-\mathbf{a}^T\mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{A}^{-1}\mathbf{a} \\
&\qquad -\mathbf{b}^T\mathbf{B}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}\mathbf{b}-2\mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b}\Big)\Big] \\
&\propto \exp\Big[-0.5\Big(\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}+\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b}-\mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{a} \\
&\qquad -\mathbf{b}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{b}-2\mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b}\Big)\Big]
\end{aligned}$$

Now we use the identities

$$\begin{aligned}
\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} &= \mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}+\mathbf{B})\mathbf{A}^{-1}\mathbf{a} \\
&= \mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a}+\mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{a}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} &= \mathbf{b}^T(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}+\mathbf{B})\mathbf{B}^{-1}\mathbf{b} \\
&= \mathbf{b}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b}+\mathbf{b}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{b}
\end{aligned}$$

Substituting these into the main equation we get

$$\begin{aligned}\kappa &\propto \exp\Big[-0.5\Big(\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}+\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b}-\mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{a}\\ &\qquad -\mathbf{b}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{b}-2\mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b}\Big)\Big]\\ &\propto \exp\Big[-0.5\Big(\mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a}+\mathbf{b}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b}-2\mathbf{a}^T(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b}\Big)\Big]\\ &\propto \exp\Big[-0.5(\mathbf{a}-\mathbf{b})^T(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{a}-\mathbf{b})\Big]\end{aligned}$$

Putting this together with the constant term we have

$$\kappa = \text{Norm}_{\mathbf{a}}[\mathbf{b}, \mathbf{A}+\mathbf{B}].$$

as required.

Problem 5.10 Prove the change of variable relation. Show that

$$\text{Norm}_{\mathbf{x}}[\mathbf{A}\mathbf{y}+\mathbf{b}, \mathbf{\Sigma}] = \kappa \cdot \text{Norm}_{\mathbf{y}}[\mathbf{A}'\mathbf{x}+\mathbf{b}', \mathbf{\Sigma}'],$$

and derive expressions for κ, $\mathbf{A}'$, $\mathbf{b}'$ and $\mathbf{\Sigma}'$. Hint: write out the terms in the original exponential, extract quadratic and linear terms in $\mathbf{y}$, and complete the square.

Answer

$$\begin{aligned}\text{Norm}_{\mathbf{x}}[\mathbf{A}\mathbf{y}+\mathbf{b}, \mathbf{\Sigma}] &= \kappa_1 \exp\Big[-0.5\Big((\mathbf{x}-\mathbf{b}-\mathbf{A}\mathbf{y})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b}-\mathbf{A}\mathbf{y})\Big)\Big]\\ &= \kappa_2 \exp\Big[-0.5\Big(\mathbf{y}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{y}-2\mathbf{y}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b})\Big)\Big]\end{aligned}$$

where we have expanded the quadratic term and absorbed the term that does not depend on $\mathbf{y}$ into a new constant κ_3. Now we complete the square to make a new distribution in the variable $\mathbf{y}$

$$\begin{aligned}\text{Norm}_{\mathbf{x}}[\mathbf{A}\mathbf{y}+\mathbf{b}, \mathbf{\Sigma}] &= \kappa_2 \exp\Big[-0.5\Big(\mathbf{y}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{y}-2\mathbf{y}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b})\Big)\Big]\\ &= \kappa_3 \exp\Big[-0.5\Big(\mathbf{y}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{y}-2\mathbf{y}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b})\\ &\qquad -2(\mathbf{x}-\mathbf{b})^T\mathbf{\Sigma}^{-1}\mathbf{A}^T(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b})\Big)\Big]\\ &= \kappa_3 \exp\Big[-0.5\Big(\Big(\mathbf{y}-(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b})\Big)^T(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})\\ &\qquad \Big(\mathbf{y}-(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b})\Big)\Big)\Big]\\ &= \kappa_4 \text{Norm}_{\mathbf{y}}[(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b}), (\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}]\end{aligned}$$

as required.

Chapter 6

Learning and inference in vision

Problem 6.1 Consider the following problems.

i Determining the gender of an image of a face.
ii Determining the pose of the human body given an image of the body.
iii Determining which suit a playing card belongs to based on an image of that card.
iv Determining whether two images of faces match (face verification).
v Determining the 3D position of a point given the positions to which it projects in two cameras at different positions in the world (stereo reconstruction).

For each case, try to describe the contents of the world state $\mathbf{w}$ and the data $\mathbf{x}$. Is each discrete or continuous? If discrete, then how many values can it take? Which are regression problems and which are classification problems?

Problem 6.2 Describe a classifier that relates univariate discrete data $x \in \{1 \ldots K\}$ to a univariate discrete world state $w \in \{1 \ldots M\}$ for both discriminative and generative model types.

Problem 6.3 Describe a regression model that relates univariate binary discrete data $x \in \{0, 1\}$ to a univariate continuous world state $w \in [-\infty, \infty]$. Use a generative formulation in which $Pr(x|w)$ and $Pr(w)$ are modeled.

Problem 6.4 Describe a discriminative regression model that relates a continuous world state $w \in [0, 1]$ to univariate continuous data $x \in [-\infty, \infty]$. Hint: Base your classifier on the Beta distribution. Ensure that the constraints on the parameters are obeyed.

Problem 6.5 Find expressions for the maximum likelihood estimates of the parameters in the discriminative linear regression model (section 6.3.1). In other words find the parameters $\{\phi_0, \phi_1, \sigma^2\}$ that satisfy

$$\begin{aligned}
\hat{\phi}_0, \hat{\phi}_1, \hat{\sigma}^2 &= \underset{\phi_0,\phi_1,\sigma^2}{\text{argmax}} \left[\prod_{i=1}^{I} Pr(w_i|x_i, \phi_0, \phi_1, \sigma^2) \right] \\
&= \underset{\phi_0,\phi_1,\sigma^2}{\text{argmax}} \left[\sum_{i=1}^{I} \log \left[Pr(w_i|x_i, \phi_0, \phi_1, \sigma^2) \right] \right] \\
&= \underset{\phi_0,\phi_1,\sigma^2}{\text{argmax}} \left[\sum_{i=1}^{I} \log \left[\text{Norm}_w \left[\phi_0 + \phi_1 x, \sigma^2 \right] \right] \right],
\end{aligned}$$

where $\{w_i, x_i\}_{i=1}^{I}$ are paired training examples.

Answer

The log likelihood L that we wish to maximize can be written out as

$$L = -\frac{I}{2}\log[\sigma^2] - \frac{I}{2}\log[2\pi] - \sum_{i=1}^{I}\frac{(w_i - \phi_0 - \phi_1 x_i)^2}{2\sigma^2}.$$

To maximize this function we take the derivative with respect to each parameters and set the resulting equation to zero. This gives three equations to solve for the three parameters. We'll start with the offset parameters ϕ_0 for which the associated derivative is:

$$\frac{\partial L}{\partial \phi_0} = \sum_{i=1}^{I}\frac{w_i - \phi_0 - \phi_1 x_i}{\sigma^2},$$

which can be re-arranged to give the relation

$$\phi_0 = \frac{\sum_i w_i - \phi_1 \sum_i x_i}{I} = \mu_w - \phi_1 \mu_x,$$

where μ_w and μ_x are the mean of the world states and data point respectively. Substituting back into the original equation we have

$$L = -\frac{I}{2}\log[\sigma^2] - \frac{I}{2}\log[2\pi] - \sum_{i=1}^{I}\frac{(w_i - \mu_w + \mu_x - \phi_1 x_i)^2}{2\sigma^2}.$$

Taking the derivative with respect to $\boldsymbol{\phi}_1$ we get

$$\frac{\partial L}{\partial \phi_0} = \sum_{i=1}^{I} x_i \frac{w_i - \mu_w + \mu_x - \phi_1 x_i}{\sigma^2}.$$

We now set this equations to zero and solve for ϕ_1 which gives us the relation

$$\phi_1 = \frac{\sum_{i=1}^{I} x_i(w_i - \mu_w + \mu_x)}{\sum_{i=1}^{I} x_i^2}$$

Finally, we take the derivative of the log likelihood with respect to σ^2 which gives

$$\frac{\partial L}{\partial \sigma^2} = -\frac{I}{2\sigma^2} + \sum_{i=1}^{I}\frac{(w_i - \phi_0 - \phi_1 x_i)^2}{2\sigma^4},$$

which can be re-arranged to get

$$\sigma^2 = \sum_{i=1}^{I}\frac{(w_i - \phi_0 - \phi_1 x_i)^2}{I}$$

Collecting these results together we have:

$$\begin{aligned}\phi_1 &= \frac{\sum_{i=1}^{I} x_i(w_i - \mu_w + \mu_x)}{\sum_{i=1}^{I} x_i^2} \\ \phi_0 &= = \mu_w - \phi_1\mu_x \\ \sigma^2 &= \sum_{i=1}^{I} \frac{(w_i - \phi_0 - \phi_1 x_i)^2}{I}.\end{aligned}$$

Problem 6.6 Consider a regression model that models the joint probability $Pr(x, w)$ between the world w and the data x as

$$Pr\left(\begin{bmatrix} w_i \\ x_i \end{bmatrix}\right) = \text{Norm}_{[w_i, x_i]^T}\left[\begin{bmatrix} \mu_w \\ \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_{ww}^2 & \sigma_{xw}^2 \\ \sigma_{xw}^2 & \sigma_{xx} \end{bmatrix}\right].$$

Use the relation in section 5.5 to compute the posterior distribution $Pr(w_i|x_i)$. Show that it has the form

$$Pr(w_i|x_i) = \text{Norm}_{w_i}[\phi_0 + \phi_1 x, \sigma^2],$$

and compute expressions for ϕ_0 and ϕ_1 in terms of the training data $\{w_i, x_i\}_{i=1}^I$ by substituting in explicit maximum likelihood estimates of the parameters $\{\mu_w, \mu_x, \sigma_{ww}^2, \sigma_{xw}^2, \sigma_{xx}^2\}$.

Problem 6.7 For a two-class problem, the *decision boundary* is the locus of world values w where the posterior probability $Pr(w = 1|x)$ is equal to 0.5. In other words, it represents the boundary between regions that would be classified as $w = 0$ and $w = 1$. Consider the generative classifier from section 6.4.2. Show that with equal priors $Pr(w = 0) = Pr(w = 1) = 0.5$ points on the decision boundary (the locus of points where $Pr(w = 0|x) = Pr(w = 1|x)$) obey a constraint of the form

$$ax^2 + bx + c = 0,$$

where and $\{a, b, c\}$ are scalars. Does the shape of the decision boundary for the logistic regression model from section 6.4.1have the same form?

Answer

Using Bayes' rule we have

$$Pr(w = 1|x) = \frac{Pr(x|w = 1)Pr(w = 1)}{Pr(x|w = 0)Pr(w = 0} + Pr(x|w = 1)Pr(w = 1).$$

Substituting in $Pr(w = 1|x) = Pr(w = 0) = Pr(w = 1) = 0.5$ we can re-arrange to get the relation $Pr(x|w = 0) = Pr(x|w = 1)$ on the boundary. So, on the decision boundary

$$\text{Norm}_x[\mu_0, \sigma_0^2] = \text{Norm}_x[\mu_1, \sigma_1^2]$$

from which it follows that

$$\log\left[\text{Norm}_x[\mu_0, \sigma_0^2]\right] = \log[\text{Norm}_x\left[\mu_1, \sigma_1^2\right]].$$

Substituting in the expressions for the log normals we have

$$-\frac{1}{2}\log[2\pi] - \frac{1}{2}\log[\sigma_0^2] - \frac{(x-\mu_0)^2}{2\sigma_0^2} = -\frac{1}{2}\log[2\pi] - \frac{1}{2}\log[\sigma_1^2] - \frac{(x-\mu_1)^2}{2\sigma_1^2}$$

Cancelling out the first terms on each side and multiplying both sides by minus two gives:

$$\log[\sigma_0^2] + \frac{(x-\mu_0)^2}{\sigma_0^2} = \log[\sigma_1^2] + \frac{(x-\mu_1)^2}{\sigma_1^2}.$$

Taking the terms of the left hand side to the right hand side gives

$$\log[\sigma_0^2] + \frac{(x-\mu_0)^2}{\sigma_0^2} - \log[\sigma_1^2] - \frac{(x-\mu_1)^2}{\sigma_1^2} = 0,$$

which can then be simplified to

$$\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)x^2 - 2\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_0}{\sigma_0^2}\right)x + \left(\frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2} + \log[\sigma_0^2] - \log[\sigma_1^2]\right) = 0,$$

which clearly is of the form $ax^2 + bx + c = 0$ as required.

Problem 6.8 Consider a generative classification model for 1D data with likelihood terms

$$\begin{aligned} Pr(x_i|w_i=0) &= \text{Norm}_{x_i}\left[0, \sigma^2\right] \\ Pr(x_i|w_i=1) &= \text{Norm}_{x_i}\left[0, 1.5\sigma^2\right]., \end{aligned}$$

What is the decision boundary for this classifier with equal priors $Pr(w=0) = Pr(w=1) = 0.5$? Develop a discriminative classifier that can produce the same decision boundary. Hint: base your classifier on a quadratic rather than a linear function.

Problem 6.9 Consider a generative binary classifier for multivariate data based on multivariate normal likelihood terms

$$\begin{aligned} Pr(\mathbf{x}_i|w_i=0) &= \text{Norm}_{\mathbf{x}_i}\left[\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right] \\ Pr(\mathbf{x}_i|w_i=1) &= \text{Norm}_{\mathbf{x}_i}\left[\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right] \end{aligned}$$

and a discriminative classifier based on logistic regression for the same data

$$Pr(w_i|\mathbf{x}_i) = \text{Bern}_{w_i}\left[\text{sig}[\phi_0 + \boldsymbol{\phi}^T\mathbf{x}_i]\right].$$

where there is one entry in the gradient vector $\boldsymbol{\phi}$ for each entry of $\mathbf{x}_i$.

How many parameters does each model have as a function of the dimensionality of $\mathbf{x}_i$? What are the relative advantages and disadvantages of each model as the dimensionality increases?

Problem 6.10 One of the problems with the background subtraction method described is that it erroneously classifies shadows as foreground. Describe a model that could be used to classify pixels into three categories (foreground, background, and shadow).

Chapter 7

Modeling complex data densities

Problem 7.1 Consider a computer vision system for machine inspection of oranges in which the goal is to tell if the orange is ripe. For each image we separate the orange from the background and calculate the average color of the pixels, which we describe as a 3×1 vector $\mathbf{x}$. We are given training pairs $\{\mathbf{x}_i, w_i\}$ of these vectors, each with an associated binary variable $w \in \{0, 1\}$ that indicates that this training example was unripe ($w = 0$) or ripe ($w = 1$). Describe how to build a generative classifier that could classify new examples $\mathbf{x}^*$ as being ripe or unripe.

Problem 7.2 It turns out that a small subset of the training labels w_i in the previous example were wrong. How could you modify your classifier to cope with this situation?

Problem 7.3 Derive the M-step equations for the mixtures of Gaussians model (equation 7.19).

Answer

From equation 7.18, the criterion that we wish to maximize is

$$L = \sum_{i=1}^{I}\sum_{k=1}^{K} r_{ik} \log\left[\lambda_k \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]\right] + \nu\left(\sum_{k=1}^{K}\lambda_k - 1\right),$$

where ν is a Lagrange multiplier that has been added to enforce the constraint that the sum of the Gaussian weights $\{\lambda_k\}_{k=1}^{K}$ is one.

Substituting in the expression for the normal distribution we have

$$\begin{aligned} L &= \sum_{i=1}^{I}\sum_{k=1}^{K} r_{ik}\left(\log[\lambda_k] - \frac{D}{2}\log 2\pi - \frac{1}{2}\log[|\boldsymbol{\Sigma}_k|] - \frac{(\mathbf{x}_i - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)}{2}\right) \\ &\quad + \nu\left(\sum_{k=1}^{K}\lambda_k - 1\right), \end{aligned}$$

Now we take derivatives with respect to $\boldsymbol{\mu}_k$, making use of the relations in Appendix C.6:

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{I} r_{ik} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k).$$

Multiplying out the terms and setting the result to zero, we get the relation

$$\boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^{I} r_{ik} \mathbf{x}_i - \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \sum_{i=1}^{I} r_{ik} = 0$$

We now multiply both sides by $\boldsymbol{\Sigma}_k$ and re-arrange to get

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{I} r_{ik} \mathbf{x}_i}{\sum_{i=1}^{I} r_{ik}},$$

as required.

Now we take derivatives with respect to $\boldsymbol{\Sigma}_k$, again making use of the relations in Appendix C.6:

$$\frac{\partial L}{\partial \boldsymbol{\Sigma}_k} = -\frac{1}{2} \sum_{i=1}^{I} r_{ik} \boldsymbol{\Sigma}_{\mathbf{k}}{}^{-T} + \sum_{i=1}^{I} r_{ik} \frac{\boldsymbol{\Sigma}^{-2} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{2}$$

Now we note that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$ because the covariance is always symmetric, multiply through by $\boldsymbol{\Sigma}^2/2$ and set the resulting expression to zero to get the result

$$-\sum_{i=1}^{I} r_{ik} \boldsymbol{\Sigma}_{\mathbf{k}} + \sum_{i=1}^{I} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T = 0,$$

which can be re-arranged to give

$$\boldsymbol{\Sigma}_{\mathbf{k}} = \frac{\sum_{i=1}^{I} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{I} r_{ik}},$$

as required.

Finally, we take the derivative with respect to λ_k to get

$$\frac{\partial L}{\partial \lambda_k} = \sum_{i=1}^{I} \frac{r_{ik}}{\lambda_k} + \nu$$

Setting this to zero and re-arranging, we get the relation

$$\sum_{i=1}^{I} r_{ik} + \nu \lambda_k = 0$$

Now we sum all K equations of this type to get

$$\sum_{k=1}^{K} \sum_{i=1}^{I} r_{ik} + \nu \sum_{k=1}^{K} \lambda_k = 0.$$

Noting that $\sum_k \lambda_k = 1$, we get an expression for ν:

$$\nu = -\sum_{k=1}^{K}\sum_{i=1}^{I} r_{ik}.$$

Now we substitute this back into the original expression for the derivative with respect to λ_k to give

$$\frac{\partial L}{\partial \lambda_k} = \sum_{i=1}^{I}\frac{r_{ik}}{\lambda_k} - \sum_{k=1}^{K}\sum_{i=1}^{I} r_{ik}.$$

Setting this to zero and re-arranging gives the relation

$$\lambda_k = \frac{\sum_{i=1}^{I} r_{ik}}{\sum_{k=1}^{K}\sum_{i=1}^{I} r_{ik}},$$

as required.

Problem 7.4 Consider modeling some univariate continuous visual data $x \in [0,1]$ using a *mixture of beta distributions*. Write down an equation for this model. Describe in words what will occur in (i) the E-step and (ii) the M-step.

Problem 7.5 Prove that the student t-distribution over x is the marginalization with respect to h of the joint distribution $Pr(x,h)$ between x and a hidden variable h where

$$\text{Stud}_x[\mu,\sigma^2,\nu] = \int \text{Norm}_x[\mu,\sigma^2/h]\text{Gam}_h[\nu/2,\nu/2]dh.$$

Answer

Substituting in the appropriate terms, we get:

$$\begin{aligned}
\text{Stud}_x[\mu,\sigma^2,\nu] &= \int_0^\infty \text{Norm}_x[\mu,\sigma^2/h]\cdot\text{Gam}_h[\nu/2,\nu/2]dh \\
&= \int_0^\infty \frac{h^{1/2}}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}h\right]\cdot\frac{\nu^{\nu/2}}{2^{\nu/2}\Gamma[\nu/2]}\exp\left[-\frac{\nu}{2}h\right]h^{\nu/2-1}dh \\
&= \frac{\nu^{\nu/2}}{2^{\nu/2}\Gamma[\nu/2]\sqrt{2\pi\sigma^2}}\int_0^\infty \exp\left[-\left(\frac{\nu}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right)h\right]h^{(\nu-1)/2}dh.
\end{aligned}$$

The integral in this expression has the form

$$\int_0^\infty \exp[-ah]h^b dh = \left[\frac{b!}{a^{b+1}}\exp[-ah]\right]_0^\infty = \frac{b!}{a^{b+1}}$$

where this was solved by using integration by parts b times in a row.

Applying this result to our problem we get

$$\begin{aligned}
\text{Stud}_x[\mu,\sigma^2,\nu] &= \frac{\nu^{\nu/2}}{2^{\nu/2}\Gamma\left[\frac{\nu}{2}\right]\sqrt{2\pi\sigma^2}}\left(\frac{\nu}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right)^{-(\nu+1)/2}((\nu-1)/2)! \\
&= \frac{\nu^{\nu/2}\Gamma[\frac{\nu+1}{2}]}{2^{\nu/2}\Gamma\left[\frac{\nu}{2}\right]\sqrt{2\pi\sigma^2}}\left(\frac{\nu}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right)^{-(\nu+1)/2} \\
&= \frac{\nu^{\nu/2}\Gamma[\frac{\nu+1}{2}]}{2^{\nu/2}\Gamma\left[\frac{\nu}{2}\right]\sqrt{2\pi\sigma^2}}\left(\frac{\nu}{2}\right)^{-(\nu+1)/2}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} \\
&= \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\sqrt{\nu\pi\sigma^2}\Gamma\left[\frac{\nu}{2}\right]}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}},
\end{aligned}$$

as required.

Problem 7.6 Show that the peak of the Gamma distribution $\text{Gam}_z[\alpha,\beta]$ is at

$$\hat{z} = \frac{\alpha-1}{\beta}.$$

Problem 7.7 Show that the Gamma distribution is conjugate to the inverse scaling factor of the variance in a normal distribution so that

$$\text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu},\boldsymbol{\Sigma}/h_i]\text{Gam}_{h_i}[\nu/2,\nu/2] = \frac{\nu^{\nu/2}}{2^{\nu/2}\Gamma[\nu/2]}\exp\left[-\frac{\nu}{2}h\right]h^{\nu/2-1},$$

and find the constant of proportionality κ and the new parameters α and β.

Answer

$$\begin{aligned}
&\text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu},\boldsymbol{\Sigma}/h_i]\cdot\text{Gam}_{h_i}[\nu/2,\nu/2] \\
&= \frac{h^{D/2}}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-0.5h(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\cdot\frac{\nu^{\nu/2}}{\Gamma[\nu/2]}\exp\left[-\frac{\nu}{2}h\right]h^{\nu/2-1} \\
&= \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\frac{\nu^{\nu/2}}{\Gamma[\nu/2]}\exp\left[-\frac{\nu+(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}h\right]h^{\nu/2+D/2-1}
\end{aligned}$$

This has the general form of a Gamma distribution but to make it complete we multiply and divide by the appropriate constants, which gives us

$$\text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu},\boldsymbol{\Sigma}/h_i]\cdot\text{Gam}_{h_i}[\nu/2,\nu/2] = \kappa\text{Beta}_h[\tilde{\alpha},\tilde{\beta}],$$

where

$$\begin{aligned}
\kappa &= \frac{\nu^{\nu/2}\Gamma[\tilde{\alpha}]}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}\tilde{\beta}^{\tilde{\alpha}}} \\
\tilde{\alpha} &= \frac{\nu+(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2} \\
\tilde{\beta} &= \nu/2+D/2
\end{aligned}$$

$$\begin{aligned}&\text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}, \boldsymbol{\Sigma}/h_i] \cdot \text{Gam}_{h_i}[\nu/2, \nu/2] \\ &= \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \frac{\nu^{\nu/2}}{\Gamma[\nu/2]} \text{Beta}_h\left[\frac{\nu + (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2},\right]\end{aligned}$$

Problem 7.8 The model for factor analysis can be written as

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{h}_i + \boldsymbol{\epsilon}_i,$$

where $\mathbf{h}_i$ is distributed normally with mean zero and identity covariance and $\boldsymbol{\epsilon}_i$ is distributed normally with mean zero and covariance $\boldsymbol{\Sigma}$. Determine expressions for

1. $E[\mathbf{x}_i]$,
2. $E[(\mathbf{x}_i - E[\mathbf{x}_i])E[\mathbf{x}_i])^T]$.

Answer

i) Applying the four expectation rules from Section 2.7, we have

$$\begin{aligned}E[\mathbf{x}_i] &= E[\boldsymbol{\mu}] + E[\boldsymbol{\Phi}\mathbf{h}_i] + \boldsymbol{\epsilon}_i \\ &= \boldsymbol{\mu} + \boldsymbol{\Phi}E[\mathbf{h}_i] + 0 \\ &= \boldsymbol{\mu} + \boldsymbol{\Phi}\cdot 0 + 0 \\ &= \boldsymbol{\mu}.\end{aligned}$$

ii) Applying the expectation rules to the second case, we have

$$\begin{aligned}E[(\mathbf{x}_i - E[\mathbf{x}_i])^2] &= E\left[(\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{h}_i + \boldsymbol{\epsilon}_i - \boldsymbol{\mu})(\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{h}_i + \boldsymbol{\epsilon}_i - \boldsymbol{\mu})^T\right] \\ &= E\left[(\boldsymbol{\Phi}\mathbf{h}_i + \boldsymbol{\epsilon}_i)(\boldsymbol{\Phi}\mathbf{h}_i + \boldsymbol{\epsilon}_i)^T\right] \\ &= E\left[\boldsymbol{\Phi}\mathbf{h}_i\mathbf{h}_i^T\boldsymbol{\Phi}^T + 2\boldsymbol{\Phi}\mathbf{h}_i\boldsymbol{\epsilon}_i^T + \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i\right] \\ &= E\left[\boldsymbol{\Phi}\mathbf{h}_i\mathbf{h}_i^T\boldsymbol{\Phi}^T\right] + E\left[2\boldsymbol{\Phi}\mathbf{h}_i\boldsymbol{\epsilon}_i^T\right] + E\left[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i\right] \\ &= \boldsymbol{\Phi}\boldsymbol{\Phi}^T E\left[\mathbf{h}_i\mathbf{h}_i^T\right] + 2\boldsymbol{\Phi}E\left[\mathbf{h}_i\right]E\left[\boldsymbol{\epsilon}_i^T\right] + E\left[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i\right] \\ &= \boldsymbol{\Phi}\boldsymbol{\Phi}^T\cdot\mathbf{I} + 2\boldsymbol{\Phi}\cdot\mathbf{0} + \boldsymbol{\Sigma} \\ &= \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Sigma}\end{aligned}$$

So, factor analysis model has a mean of $\boldsymbol{\mu}$ and a covariance of $\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Sigma}$ which is exactly what we claimed when we marginalized out the hidden variable:

$$\begin{aligned}Pr(\mathbf{x}) &= \int Pr(\mathbf{x}_i|\mathbf{h}_i)Pr(\mathbf{h}_i)d\mathbf{h}_i \\ &= \int \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{h}_i, \boldsymbol{\Sigma}]\text{Norm}_{\mathbf{h}_i}[\mathbf{0}, \mathbf{I}]d\mathbf{h}_i \\ &= \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Sigma}].\end{aligned}$$

Problem 7.9 Derive the E-step for factor analysis (equation 7.34).

Answer

We begin by applying the change in variables relation (section 5.7) which gives us

$$\begin{aligned}\hat{q}(\mathbf{h}_i) &\propto \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}+\boldsymbol{\Phi}\mathbf{h}_i, \boldsymbol{\Sigma}]\text{Norm}_{\mathbf{h}_i}[\mathbf{0}, \mathbf{I}] \\ &= \text{Norm}_{\mathbf{h}_i}[(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}), (\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi})^{-1}]\text{Norm}_{\mathbf{h}_i}[\mathbf{0}, \mathbf{I}]\end{aligned}$$

Now we apply the product of two normal distributions rule (section 5.6) which gives us

$$\begin{aligned}\hat{q}(\mathbf{h}_i) &\propto \text{Norm}_{\mathbf{h}_i}[(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}), (\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi})^{-1}]\text{Norm}_{\mathbf{h}_i}[\mathbf{0}, \mathbf{I}] \\ &= \text{Norm}_{\mathbf{h}_i}[(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}+\mathbf{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}), (\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}+\mathbf{I})^{-1}]\end{aligned}$$

as required. Note that the constant must be one to make the posterior a valid probability distribution.

Problem 7.10 Derive the M-step for factor analysis (equation 7.38).

Answer

The log likelihood is given by

$$\begin{aligned}L &= -\frac{1}{2}E\left[\sum_{i=1}^{I}\left(2D\log 2\pi + 2\log[|\boldsymbol{\Sigma}|] + (\mathbf{x}_i-\boldsymbol{\mu}-\Phi\mathbf{h}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}-\Phi\mathbf{h}_i)\right)\right] \\ &= -\frac{1}{2}E\left[\sum_{i=1}^{I}\left(2D\log 2\pi + 2\log[|\boldsymbol{\Sigma}|] + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i + \mathbf{h}_i\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\mathbf{h}_i\right.\right. \\ &\qquad\left.\left. -2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i - 2\mu^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\mathbf{h}_i - 2\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\mathbf{h}_i\right)\right]\end{aligned}$$

Taking the derivative with respect to $\boldsymbol{\mu}$ we get

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\mu}} &= -\frac{1}{2}E\left[\sum_{i=1}^{I}\left(2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\Sigma}^{-1}\mathbf{x}_i - 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\mathbf{h}_i\right)\right] \\ &= \sum_{i=1}^{I}\left(-\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + 2\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right) = 0\end{aligned}$$

which we can re-arrange to get

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{I}\mathbf{x}_i}{I}$$

Taking the derivative with respect to $\boldsymbol{\Phi}$ we get

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\Phi}} &= -\frac{1}{2}E\left[\sum_{i=1}^{I}\left(2\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\mathbf{h}_i\mathbf{h}_i^T - 2\boldsymbol{\Sigma}^{-1}\mu\mathbf{h}_i^T - 2\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\mathbf{h}_i^T\right)\right] \\ &= \sum_{i=1}^{I}\left(2\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}E[\mathbf{h}_i\mathbf{h}_i^T] - 2\boldsymbol{\Sigma}^{-1}\mu E[\mathbf{h}_i^T] - 2\boldsymbol{\Sigma}^{-1}\mathbf{x}_i E[\mathbf{h}_i^T]\right) = 0\end{aligned}$$

This can be re-arranged to get

$$\hat{\boldsymbol{\Phi}} = \left(\sum_{i=1}^{I}(\mathbf{x}_i - \boldsymbol{\mu})E[\mathbf{h}_i]^T\right)\left(\sum_{i=1}^{I}E[\mathbf{h}_i\mathbf{h}_i^T]\right)^{-1}$$

Chapter 8

Regression models

Problem 8.1 Consider a regression problem where the world state w is known to be positive. To cope with this we could construct a regression model in which the world state is modeled as a gamma distribution. We could constrain both parameters α, β of the gamma distribution to be the same so that $\alpha = \beta$ and make them a function of the data $\mathbf{x}$. Describe a maximum likelihood approach to fitting this model.

Problem 8.2 Consider a robust regression problem based on the t-distribution rather than the normal distribution. Define this model precisely in mathematical terms and sketch out a maximum likelihood approach to fitting the parameters.

Problem 8.3 Prove that the maximum likelihood solution for the gradient in the linear regression model is

$$\hat{\boldsymbol{\phi}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{w}.$$

Answer

To fit the parameters, we optimize the criterion

$$\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\text{argmax}}\left[\log[\text{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi}, \sigma^2\mathbf{I}]\right] \\
&= \underset{\boldsymbol{\phi}}{\text{argmax}}\left[-\frac{D}{2}\log[\sigma^2] - \frac{D}{2}\log[2\pi] - \frac{1}{2\sigma^2}(\mathbf{w}-\mathbf{X}^T\boldsymbol{\phi})^T(\mathbf{w}-\mathbf{X}^T\boldsymbol{\phi})\right] \\
&= \underset{\boldsymbol{\phi}}{\text{argmax}}\left[-\frac{1}{2\sigma^2}(\mathbf{w}-\mathbf{X}^T\boldsymbol{\phi})^T(\mathbf{w}-\mathbf{X}^T\boldsymbol{\phi})\right] \\
&= \underset{\boldsymbol{\phi}}{\text{argmax}}\left[-\frac{1}{2\sigma^2}\left(\mathbf{w}^T\mathbf{w} + \boldsymbol{\phi}^T\mathbf{X}\mathbf{X}^T\boldsymbol{\phi} - 2\mathbf{w}\mathbf{X}^T\boldsymbol{\phi}\right)\right] \\
&= \underset{\boldsymbol{\phi}}{\text{argmax}}\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\phi}^T\mathbf{X}\mathbf{X}^T\boldsymbol{\phi} - 2\mathbf{w}\mathbf{X}^T\boldsymbol{\phi}\right)\right] \\
&= \underset{\boldsymbol{\phi}}{\text{argmax}}\left[-\boldsymbol{\phi}^T\mathbf{X}\mathbf{X}^T\boldsymbol{\phi} + 2\mathbf{w}\mathbf{X}^T\boldsymbol{\phi}\right]
\end{aligned}$$

Denoting this criterion by L and then taking the derivative with respect to $\boldsymbol{\phi}$ and setting the result to zero we get

$$\frac{\partial L}{\partial \phi} = -2\mathbf{X}\mathbf{X}^T\boldsymbol{\phi} + 2\mathbf{X}\mathbf{w} = 0$$

Re-arranging this equation gives us the desired relation:

$$\hat{\boldsymbol{\phi}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{w}.$$

Problem 8.4 For the Bayesian linear regression model (section 8.2), show that the posterior distribution over the parameters $\boldsymbol{\phi}$ is given by

$$Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) = \text{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right],$$

where

$$\mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}.$$

Answer

The posterior is computed via Bayes rule so

$$\begin{aligned} Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) &\propto Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\phi})Pr(\boldsymbol{\phi}) \\ &= \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi}, \sigma^2\mathbf{I}]\text{Norm}_{\boldsymbol{\phi}}[\mathbf{0}, \sigma_p^2\mathbf{I}] \end{aligned}$$

Using the change of variables relation from section 5.7 we see that

$$Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) \propto \text{Norm}_{\boldsymbol{\phi}}[(\sigma^{-2}\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\sigma^{-2}\mathbf{w}, (\sigma^{-2}\mathbf{X}\mathbf{X}^T)^{-1}]\text{Norm}_{\boldsymbol{\phi}}[\mathbf{0}, \sigma_p^2\mathbf{I}]$$

Using the product of tow normals rule form section 5.6 we get

$$Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) \propto \text{Norm}_{\boldsymbol{\phi}}[(\sigma^{-2}\mathbf{X}\mathbf{X}^T + \sigma_p^{-2}\mathbf{I})^{-1}\mathbf{X}\sigma^{-2}\mathbf{w}, (\sigma^{-2}\mathbf{X}\mathbf{X}^T)^{-1}]$$

It is now easy to see that this has the desired form. Note that the constant of proportionality must be one since the posterior distribution must be a valid probability distribution.

Problem 8.5 For the Bayesian linear regression model (section 8.2) show that the predictive distribution for a new data example $\mathbf{x}^*$ is given by

$$Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) = \text{Norm}_{w^*}\left[\frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{x}^* + \sigma^2\right].$$

Answer

$$\begin{aligned} Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*, \boldsymbol{\phi}) Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) d\boldsymbol{\phi} \\ &= \int \text{Norm}_{w^*}[\boldsymbol{\phi}^T\mathbf{x}^*, \sigma^2] \text{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right] d\boldsymbol{\phi} \end{aligned}$$

It is quite arduous to prove this the straightforward way, so we'll take a shortcut. First, we note that the predictive distribution will be normal. The normal distribution is self-conjugate with respect to its own mean and after reformulating the two terms in the integral as a single distribution in $\boldsymbol{\phi}$, the remaining constant will also be normal.

Our goal now is to compute the mean and covariance of this normal. To do this, we'll write both distributions from the original integral as generative equations:

$$\begin{aligned} w^* &= \mathbf{x}^{*T}\boldsymbol{\phi} + \epsilon_i \\ \boldsymbol{\phi} &= \frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w} + \boldsymbol{\alpha}_i \end{aligned}$$

where ϵ_i is a noise term with mean 0 and variance σ^2 and $\boldsymbol{\alpha}$ is a second noise term with mean $\mathbf{0}$ and covariance $\mathbf{A}^{-1}$.

Now we compute the expected value of w^* which is

$$\begin{aligned} E[w^*] &= \mathbf{x}^{*T}E[\boldsymbol{\phi}] + E[\epsilon_i] \\ &= \mathbf{x}^{*T}E[\boldsymbol{\phi}] \\ &= \frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{X}\mathbf{w} \end{aligned}$$

which is indeed the claimed mean of the predictive distribution.

Now we compute the second moment about zero

$$\begin{aligned} E[w^{*2}] &= \mathbf{x}^{*T}E[\boldsymbol{\phi}\boldsymbol{\phi}^T]\mathbf{x}^* + E[\epsilon_i\epsilon_i^T] \\ &= \mathbf{x}^{*T}E[\boldsymbol{\phi}\boldsymbol{\phi}^T]\mathbf{x}^* + \sigma^2 \end{aligned}$$

We note that

$$\begin{aligned} E[\boldsymbol{\phi}\boldsymbol{\phi}^T] &= E[(\boldsymbol{\phi} - E[\boldsymbol{\phi}])^T(\boldsymbol{\phi} - E[\boldsymbol{\phi}])] + E[\boldsymbol{\phi}]E[\boldsymbol{\phi}^T] \\ &= \mathbf{A}^{-1} + \frac{1}{\sigma^4}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T\mathbf{A}^{-1} \end{aligned}$$

and so the covariance is given by

$$\begin{aligned} E[(w - E[w])^2] &= E[w^{*2}] - E[w^*]^2 \\ &= \mathbf{x}^{*T}E[\boldsymbol{\phi}\boldsymbol{\phi}^T]\mathbf{x}^* + \sigma^2 - \frac{1}{\sigma^4}\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T\mathbf{A}^{-1}\mathbf{x}^* \\ &= \mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{x}^* + \sigma^2, \end{aligned}$$

as required.

Problem 8.6 Use the matrix inversion lemma (appendix C.8.4) to show that

$$\mathbf{A}^{-1} = \left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}_D\right)^{-1} = \sigma_p^2\mathbf{I}_D - \sigma_p^2\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}_I\right)^{-1}\mathbf{X}^T.$$

Problem 8.7 Compute the derivative of the marginal likelihood

$$Pr(\mathbf{w}|\mathbf{X},\sigma^2) \quad = \quad \text{Norm}_{\mathbf{w}}[\mathbf{0}, \sigma_p^2\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{I}],$$

with respect to the variance parameter σ^2.

Problem 8.8
Compute a closed form expression for the approximated t-distribution used to impose sparseness.

$$q(h) = \max_h \left[\text{Norm}_\phi[0, h^{-1}]\text{Gam}_h[\nu/2, \nu/2]\right].$$

Plot this function for $\nu = 2$. Plot the 2D function $[h_1, h_2] = q(h_1)q(h_2)$ for $\nu = 2$.

Problem 8.9 Describe maximum likelihood learning and inference algorithms for a non-linear regression model based on polynomials where

$$Pr(w|x) = \text{Norm}_w[\phi_0 + \phi_1 x + \phi_2 x^2 + \phi_3 x^3, \sigma^2].$$

Problem 8.10 I wish to learn a linear regression model in which I predict the world w from I examples of $D \times 1$ data $\mathbf{x}$ using the maximum likelihood method. If $I > D$ is it more efficient to use the dual parameterization or the original linear regression model?

Problem 8.11 Show that the maximum likelihood estimate for the parameters $\boldsymbol{\psi}$ in the dual linear regression model (section 8.7) is given by

$$\hat{\boldsymbol{\psi}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w}.$$

Chapter 9

Classification models

Problem 9.1 The logistic sigmoid function is defined as

$$\text{sig}[a] = \frac{1}{1+\exp[-a]}.$$

Show that (i) $\text{sig}[-\infty] = 0$, (ii) $\text{sig}[0] = 0.5$, (iii) $\text{sig}[\infty] = 1$.

Problem 9.2 Show that the derivative of the log posterior probability for the logistic regression model

$$L = \sum_{i=1}^{I} w_i \log\left[\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right] + \sum_{i=1}^{I}(1-w_i)\log\left[\frac{\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\right]$$

with respect to the parameters $\boldsymbol{\phi}$ is given by

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = -\sum_{i=1}^{I}\left(\text{sig}[a_i] - w_i\right)\mathbf{x}_i.$$

Answer

First, we'll simplify the criterion by sorting out the log terms

$$\begin{aligned}
L &= \sum_{i=1}^{I}\left(-w_i\log\left[1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]\right] + (1-w_i)\log\left[\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]\right] - (1-w_i)\log\left[1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]\right]\right) \\
&= \sum_{i=1}^{I}\left((1-w_i)\log\left[\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]\right] - \log\left[1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]\right]\right) \\
&= \sum_{i=1}^{I}\left(-(1-w_i)\boldsymbol{\phi}^T\mathbf{x}_i - \log\left[1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]\right]\right)
\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\phi}} &= \sum_{i=1}^{I}\left(-(1-w_i)\mathbf{x}_i - \frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\cdot\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]\cdot-\mathbf{x}_i\right)\\ &= \sum_{i=1}^{I}\left((1-w_i)\mathbf{x}_i + (1-\text{sig}[a_i])\mathbf{x}_i\right)\\ &= -\sum_{i=1}^{I}\left(\text{sig}[a_i]-w_i\right)\mathbf{x}_i\end{aligned}$$

as required.

Problem 9.3 Show that the second derivative of the log likelihood of the logistic regression model is given by

$$\frac{\partial^2 L}{\partial \boldsymbol{\phi}^2} = -\sum_{i=1}^{I}\text{sig}[a_i](1-\text{sig}[a_i])\mathbf{x}_i\mathbf{x}_i^T.$$

Answer

$$\begin{aligned}\frac{\partial^2 L}{\partial \boldsymbol{\phi}^2} &= \frac{\partial}{\partial \boldsymbol{\phi}}\left[-\sum_{i=1}^{I}\left(\text{sig}[a_i]-w_i\right)\mathbf{x}_i\right]\\ &= -\sum_{i=1}^{I}\frac{1}{(1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i])^2}\cdot\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]\mathbf{x}_i\mathbf{x}_i^T\\ &= -\sum_{i=1}^{I}\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\cdot\frac{\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}_i]}\mathbf{x}_i\mathbf{x}_i^T\\ &= -\sum_{i=1}^{I}\text{sig}[a_i](1-\text{sig}[a_i])\mathbf{x}_i\mathbf{x}_i^T,\end{aligned}$$

---

as required.

Problem 9.4 Consider fitting a logistic regression data to 1D data x where the two classes are perfectly separable. For example, perhaps all the data x where the world state $w=0$ takes values less than 0 and all the data x where the world state is $w=1$ takes values greater than 1. Hence it is possible to classify the training data perfectly. What will happen to the parameters of the model during learning? How could you rectify this problem?

Problem 9.5 Compute the Laplace approximation to a beta distribution with parameters $\alpha = 1.0$, $\beta = 1.0$.

Problem 9.6 Show that the Laplace approximation to a univariate normal distribution with mean μ and variance σ^2 is the normal distribution itself.

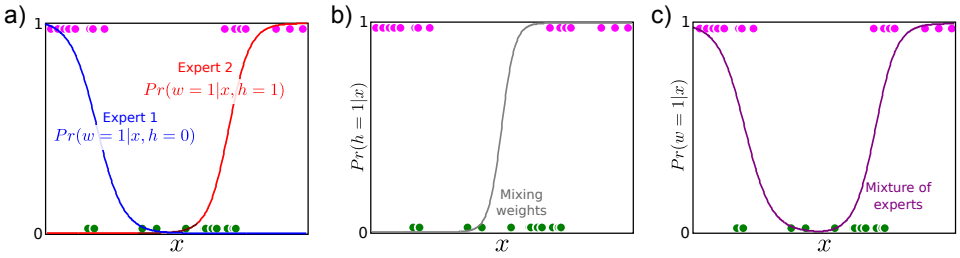

Figure 9.1 Mixture of two experts model for 1D data. Pink circles indicate positive examples. Green circles indicate negative examples. a) Two expert is specialized to model the left and right sides of the data, respectively. b) The mixing weights change as a function of the data. c) The final output of the model is mixture of the two constituent experts and fits the data well.

Problem 9.7 Show that the second derivative of the logarithm $L = \log[\text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]]$ of a normal distribution evaluated at the mean $\boldsymbol{\mu}$ is given by

$$\left|\frac{\partial L}{\partial \phi^2}\right|_{\boldsymbol{\mu}} = -\boldsymbol{\Sigma}^{-1}.$$

Problem 9.8 Devise a method to choose the scale parameter λ_0 in the radial basis function in kernel logistic regression (equation 9.33).

Problem 9.9

A *mixture of experts* (Jordan and Jacobs, 1994) divides space into different regions, each of which receives specialized attention (figure 9.1) . For example, we could describe the data as a mixture of logistic classifiers so that

$$Pr(w_i|\mathbf{x}_i) = \sum_{k=1}^{K} \lambda_k[\mathbf{x}_i]\text{Bern}_{w_i}\left[\text{sig}[\boldsymbol{\phi}_k^T\mathbf{x}_i]\right].$$

Each logistic classifier is considered as an expert and the mixing weights decide the combination of experts that are applied to the data. The mixing weights, which are positive and sum to one, depend on the data $\mathbf{x}$: for a two-component model they could be based on a second logistic regression model with activation $\boldsymbol{\omega}^T\mathbf{x}$. This model can be expressed as the marginalization of a joint distribution between $\mathbf{w}_i$ and a hidden variable h_i so that

$$\sum_{k=1}^{K} \lambda_k[\mathbf{x}_i] = 1 \qquad \forall \ \mathbf{x}_i$$

For example, to divide the data into two regions we could model the values λ_k based on the outputs of a second logistic regression model with parameters $\boldsymbol{\omega}$.

$$Pr(w_i|\mathbf{x}_i) = \sum_{k=1}^{K} Pr(w_i, h_i = k|\mathbf{x}_i) = \sum_{k=1}^{K} Pr(w_i|h_i = k, \mathbf{x}_i)Pr(h_i = k|\mathbf{x}_i),$$

where

$$\begin{aligned} Pr(w_i|h_i=k,\mathbf{x}_i) &= \text{Bern}_{w_i}\left[\text{sig}[\boldsymbol{\phi}_k^T\mathbf{x}_i]\right] \\ Pr(h_i=k|\mathbf{x}_i) &= \text{Bern}_{h_i}\left[\text{sig}[\boldsymbol{\omega}^T\mathbf{x}_i]\right]. \end{aligned}$$

How does this model differ from branching logistic regression (section 9.8)? Devise a learning algorithm for this model.

Answer

The mixtures of experts is very closely related to the branching logistic regression model. In the former case the prediction is based on a linear sum of logistic regression models associated with each expert. In the latter the prediction is the result of a single logistic regression model where the activation is a linear sum of terms associated with each expert. As for the branching logistic regression model, the mixtures of experts can be generalized so the mixing weights or the experts themselves contain a nonlinear activation term. They may also be built hierarchically to form a tree structure.

The model can either be learnt by direct optimization of the log posterior probability, or using the EM algorithm. In the latter case, a hidden variable is associated with each data examples. This can be interpreted as indicating which of the logistic regression models a data example is assigned to. In the E-step we calculate a posterior distribution over the I hidden variables $\{h_i\}_{i=1}^I$ associated with each of the data examples. In the M-step we maximize the bound with respect to the parameters $\boldsymbol{\theta}=\{\boldsymbol{\phi}_0,\boldsymbol{\phi}_1,\boldsymbol{\omega}\}$.

Problem 9.10 The $\textbf{softmax}[\bullet,\bullet,\ldots,\bullet]$ function is defined to return a multivariate quantity where the k^{th} element is given by

$$s_k = \text{softmax}_k[a_1,a_2,\ldots a_K] = \frac{\exp[a_k]}{\sum_{j=1}^K \exp[a_j]}.$$

Show that $0<s_k<1$ and that $\sum_{k=1}^K s_k = 1$.

Problem 9.11 Show that the first derivative of the log-probability of the multi-class logistic regression model is given by equation 9.61.

Answer

The cost function is given by

$$\begin{aligned} L &= \sum_{i=1}^I\sum_{n=1}^K \delta[w_i-n]\log\left[\frac{\exp[\boldsymbol{\phi}_n^T\mathbf{x}_i]}{\sum_{j=1}^K\exp[\boldsymbol{\phi}_j^T\mathbf{x}_i]}\right] \\ &= \sum_{i=1}^I\left(-\log\left[\sum_{j=1}^J\exp[\boldsymbol{\phi}_j^T\mathbf{x}_i]\right]+\sum_{n=1}^N\delta[w_i-n]\boldsymbol{\phi}_n^T\mathbf{x}_i\right) \end{aligned}$$

Now we compute the derivative

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\phi}_n} &= \sum_{i=1}^{I}\left(-\frac{\exp[\boldsymbol{\phi}_n^T\mathbf{x}_i}{\sum_{j=1}^{K}\exp[\boldsymbol{\phi}_j^T\mathbf{x}_i}\mathbf{x}_i + \delta[w_i - n]\mathbf{x}\right) \\ &= -\sum_{i=1}^{I}(y_{in} - \delta[w_i - n])\mathbf{x}_i\end{aligned}$$

Problem 9.12 The classifiers in this chapter have all been based on continuous data $\mathbf{x}$. Devise a model that can distinguish between M world states $w \in \{1 \ldots M\}$ based on a discrete observation $x \in \{1 \ldots K\}$ and discuss potential learning algorithms.

Chapter 10

Graphical models

Problem 10.1 The joint probability model between variables $\{x_n\}_{n=1}^{7}$ factorizes as

$$Pr(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = Pr(x_1)Pr(x_3)Pr(x_7)Pr(x_2|x_1, x_3)Pr(x_5|x_7, x_2)Pr(x_4|x_2)Pr(x_6|x_5, x_4).$$

Draw a directed graphical model relating these variables. Which variables form the Markov blanket of variable x_2?

Problem 10.2 Write out the factorization corresponding to the directed graphical model in figure 10.1a.

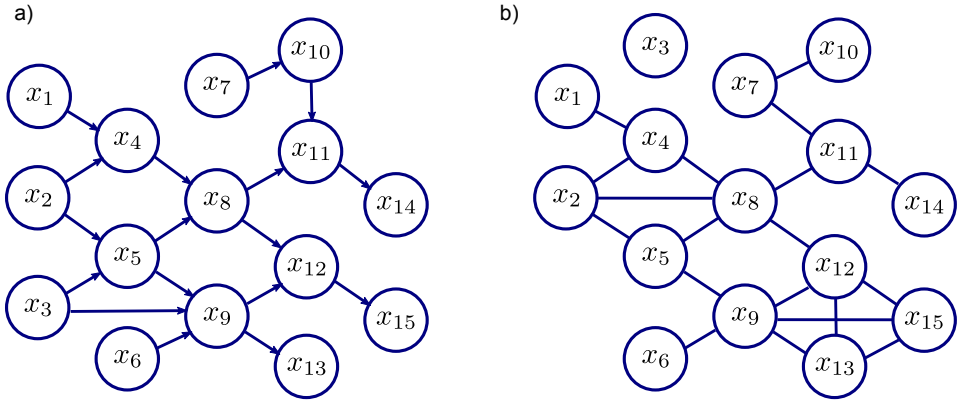

Figure 10.1 a) Graphical model for problem 10.2. b) Graphical model for problem 10.4

Problem 10.3 An undirected graphical model has the form

$$Pr(x_1 \ldots x_6) = \frac{1}{Z}\Phi_1[x_1, x_2, x_5]\Phi_2[x_2, x_3, x_4]\Phi_3[x_1 x_5]\Phi_4[x_5, x_6].$$

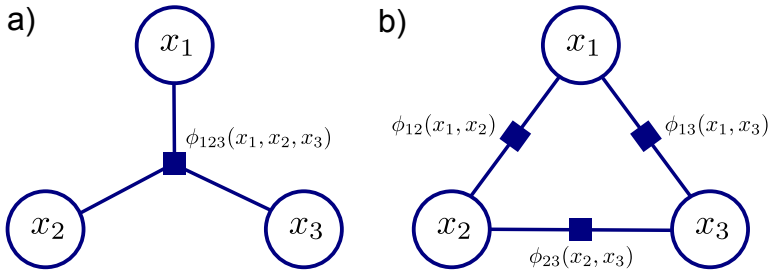

Figure 10.2 Factor graphs contain one node (square) per factor in the joint pdf as well as one node (circle) per variable. Each factor node is connected to all of the variables that belong to that factor. This type of graphical model can distinguish between the undirected graphical models a) $Pr(x_1, x_2, x_3) = \frac{1}{Z}\phi_{123}[x_1, x_2, x_3]$ and b) $Pr(x_1, x_2, x_3) = \frac{1}{Z}\phi_{12}[x_1, x_2]\phi_{23}[x_2, x_3]\phi_{13}[x_1, x_3]$.

Draw the undirected graphical model that corresponds to this factorization.

Problem 10.4 Write out the factorization corresponding to the undirected graphical model in figure 10.14b from the book.

Problem 10.5 Consider the undirected graphical model defined over binary values $\{x_i\}_{i=1}^4 \in \{0, 1\}$ defined by

$$Pr(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi(x_1, x_2)\phi(x_2, x_3)\phi(x_3, x_4)\phi(x_4, x_1),$$

where the function ϕ is defined by

$$\begin{aligned} \phi(0,0) &= 1 & \phi(1,1) &= 2 \\ \phi(0,1) &= 0.1 & \phi(1,0) &= 0.1 \end{aligned}$$

Compute the probability of each of the 16 possible states of this system.

Problem 10.6 What is the Markov blanket for each of the variables in figures 10.7 and 10.8 from the book?

Problem 10.7 Show that the stated patterns of independence and conditional independence in figure 10.7 and figure 10.8 from the book are true.

Problem 10.8 A *factor graph* is a third type of graphical model that depicts the factorization of a joint probability. As usual it contains a single node per variable, but it also contains one node per factor (usually indicated by a solid square). Each factor variable is connected to all of the variables that are contained in the associated term in the factorization by undirected links. For example, the factor node corresponding to the term $Pr(x_1|x_2, x_3)$ in a directed model would connect to all three variables x_1, x_2 and x_3. Similarly, the factor node corresponding to the term $\phi_{12}[x_1, x_2]$ in an undirected model would connect variables x_1 and x_2. Figure 10.2 shows two examples of factor graphs.

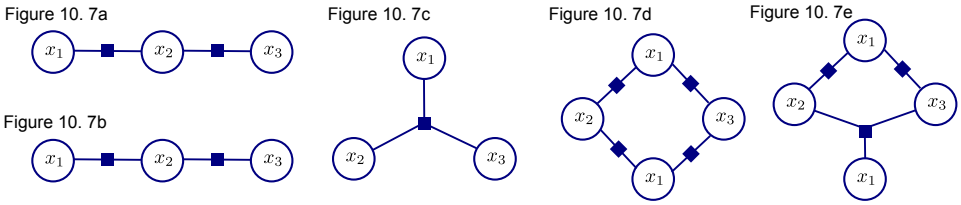

Figure 10.3 Factor graphs for problem 10.8

Draw the factor graphs corresponding to the graphical models in figures 10.7 and 10.8 from the book. You must first establish the factorized joint distribution associated with each graph.

Answer

The associated factor graphs are illustrated in figure 10.3.

Problem 10.9 What is the Markov blanket of variable w_2 in figure 10.9c from the book?

Problem 10.10 What is the Markov blanket of variable w_8 in figure 10.9e from the book?

Chapter 11

Models for chains and trees

Problem 11.1 Compute by hand the lowest possible cost for traversing the graph in figure 11.1 using the dynamic programming method.

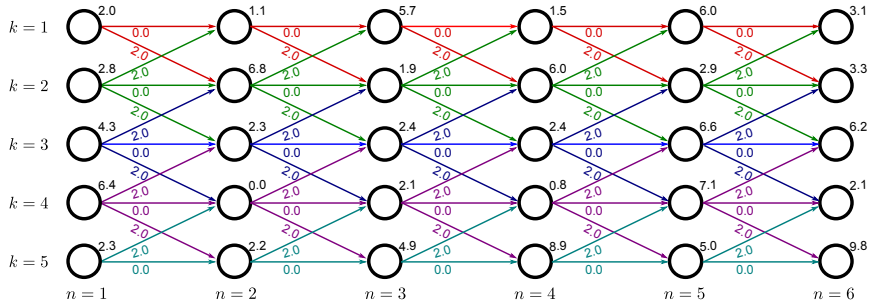

Figure 11.1 Dynamic programming example for problem 11.1.

Problem 11.2 MAP inference in chain models can also be performed by running Djikstra's algorithm on the graph in figure 11.2, starting from the node on the left hand side and terminating when we first reach the node on the right hand side. If there are N variables, each of which takes K values, what is the best and worst case complexity of the algorithm? Describe a situation where Djikstra's algorithm outperforms dynamic programming.

Problem 11.3 Consider the graphical model in figure 11.3a. Write out the cost function for MAP estimation in the form of equation 11.17. Discuss the difference between your answer and equation 11.17.

Problem 11.4 Compute the solution (minimum cost path) to the dynamic programming problem on the tree in figure 11.4.

Problem 11.5 MAP inference for the chain model can be expressed as

$$\hat{w}_N = \underset{w_N}{\text{argmax}}\left[\max_{w_1}\left[\max_{w_2}\left[\ldots \max_{w_{N-1}}\left[\sum_{n=1}^{N}\log[Pr(\mathbf{x}_n|w_n)] + \sum_{n=2}^{N}\log[Pr(w_n|w_{n-1})]\right]\ldots\right]\right]\right]$$

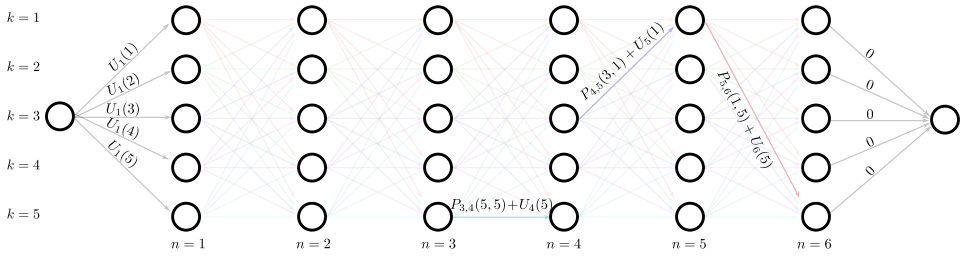

Figure 11.2 Graph construction for problem 11.2. This is the same as the dynamic programming graph (figure 11.3) except that: (i) there are two extra nodes at the start and the end of the graph. (ii) There are no vertex costs. (iii) The costs associated with the left-most edges are $U_1(k)$ and the costs associated with the right-most edges are 0. The general edge cost for passing from label a and node n to label b at node $n+1$ is given by $P_{n,n+1}(a,b) + U_{n+1}(b)$.

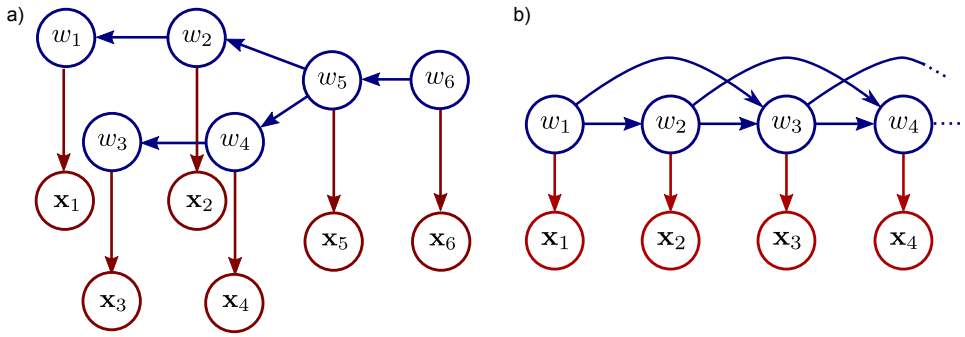

Figure 11.3 a) Graphical model for problem 11.3. b) Graphical model for problem 11.10. The unknown variables $w_3, w_4 \ldots$ in this model receive connections from the two preceding variables and so the graph contains loops.

Show that it is possible to compute this expression piecewise by moving the maximization terms through the summation sequence in a manner similar to that described in section 11.4.1.

Problem 11.6 Develop an algorithm that can compute the marginal distribution for an arbitrary variable w_n in a chain model.

Problem 11.7 Develop an algorithm that computes the joint marginal distribution of any two variables w_m and w_n in a chain model.

Problem 11.8 Consider the following two distributions over three variables x_1, x_2, and x_3:

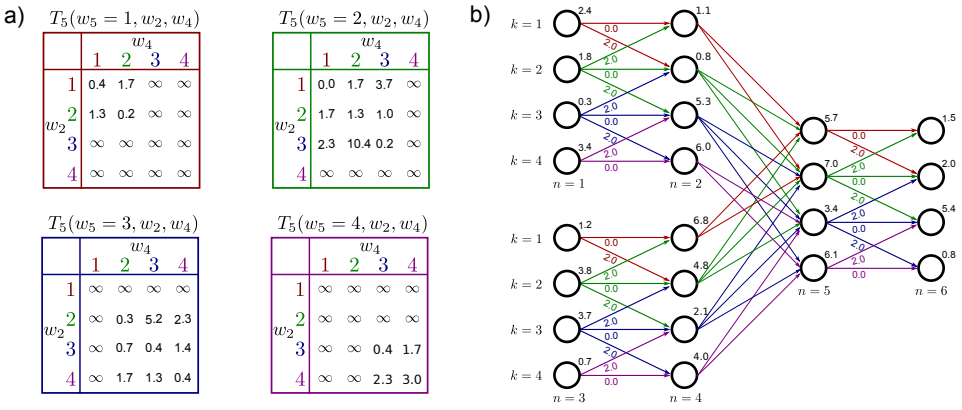

Figure 11.4 Dynamic programming example for problem 11.4.

$$\begin{aligned} Pr(x_1, x_2, x_3) &= \frac{1}{Z_1}\phi_{12}[x_1, x_2]\phi_{23}[x_2, x_3]\phi_{31}[x_3, x_1] \\ Pr(x_1, x_2, x_3) &= \frac{1}{Z_2}\phi_{123}[x_1, x_2, x_3]. \end{aligned}$$

Draw (i) an undirected model and (ii) a factor graph for each distribution. What do you conclude?

Problem 11.9 Convert each of the graphical models in figure 11.5 into the form of a factor graph. Which of the resulting factor graphs take the form of a chain?

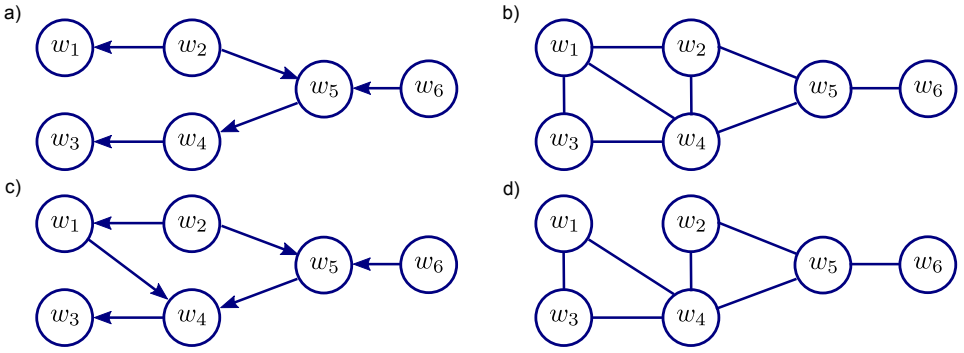

Figure 11.5 Answers for problem 11.9.

Problem 11.10 Figure 11.3b shows a chain model in which each unknown variable w depends on its two predecessors. Describe a dynamic programming approach to finding the MAP solution. (hint: you need to combine variables.) If there are N variables in the chain and each takes K values, what is the overall complexity of your algorithm?

Problem 11.11 In the stereo vision problem, the solution was very poor when the pixels are treated independently. Suggest some improvements to this method (while keeping the pixels independent).

Problem 11.12 Consider a variant on the segmentation application in which we update all of the contour positions at once. The graphical model for this problem is a loop (i.e., a chain where there is also a edge between w_N and w_1). Devise an approach to finding the exact MAP solution in this model. If there are N variables each of which can take K values, what is the complexity of your algorithm?

Chapter 12

Models for grids

Problem 12.1 Consider a Markov random field with the structure

$$Pr(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi[x_1, x_2]\phi[x_2, x_3]\phi[x_3, x_4]\phi[x_4, x_1]$$

but where the variables x_1, x_2, x_3, and x_4 are continuous and the potentials are defined as

$$\phi[a, b] = \exp\left[-(a-b)^2\right].$$

This is known as a *Gaussian Markov random field*. Show that the joint probability is a normal distribution, and find the information matrix (inverse covariance matrix).

Answer

Filling in the missing elements of the equation we have:

$$\begin{aligned}
Pr(x_1, x_2, x_3, x_4) &= \frac{1}{Z}\phi[x_1, x_2]\phi[x_2, x_3]\phi[x_3, x_4]\phi[x_4, x_1] \\
&= \frac{1}{Z}\exp\left[-(x_1-x_2)^2\right]\exp\left[-(x_2-x_3)^2\right]\exp\left[-(x_3-x_4)^2\right]\exp\left[-(x_4-x_1)^2\right] \\
&= \frac{1}{Z}\exp\left[-(x_1-x_2)^2-(x_2-x_3)^2-(x_3-x_4)^2-(x_4-x_1)^2\right] \\
&= \frac{1}{Z}\exp\left[-\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}\begin{bmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}\right]
\end{aligned}$$

Problem 12.2 Compute the MAP solution to the three-pixel graph cut problem in figure 12.1 by (i) computing the cost of all eight possible solutions explicitly and finding the one with the minimum cost (ii) running the augmenting paths algorithm on this graph by hand and interpreting the minimum cut.

Problem 12.3 Explicitly compute the costs associated with the four possible minimum cuts of the graph in figure 12.10 from the book.

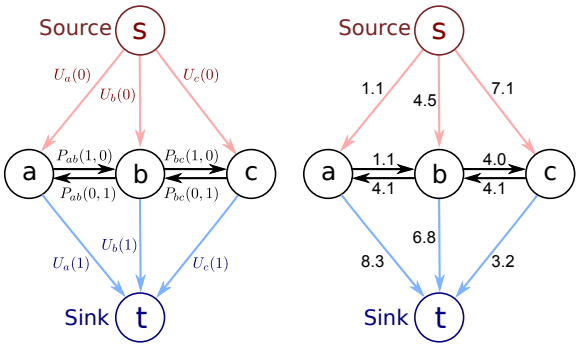

Figure 12.1 Graph for problem 12.2.

Problem 12.4 Compute the cost for each the four possible cuts of the graph in figure 12.11c from the book.

Problem 12.5 Consider the graph construction in figure 12.2a, which contains a number of constraint edges of infinite cost (capacity). There are 25 possible minimum cuts on this graph, each of which corresponds to one possible labeling of the two pixels. Write out the cost for each labeling. Which solutions have finite cost for this graph construction?

Answer

The costs are as follows:

$$\begin{aligned}
C[1,1] &= U_a(1) + U_b(1) \\
C[1,2] &= U_a(1) + U_b(1) + P_{ab}(1,2) \\
C[1,3] &= \infty \\
C[1,4] &= \infty \\
C[1,5] &= \infty
\end{aligned}$$

$$\begin{aligned}
C[2,1] &= U_a(2) + U_b(1) + P_{ab}(2,1) \\
C[2,2] &= U_a(2) + U_b(2) \\
C[2,3] &= U_a(2) + U_b(3) + P_{ab}(2,3) \\
C[2,4] &= \infty \\
C[2,5] &= \infty
\end{aligned}$$

$$\begin{aligned}
C[3,1] &= \infty \\
C[3,2] &= U_a(3) + U_b(2) + P_{ab}(3,2) \\
C[3,3] &= U_a(3) + U_b(3) \\
C[3,4] &= U_a(3) + U_b(4) + P_{ab}(3,4) \\
C[3,5] &= \infty
\end{aligned}$$

$$
\begin{aligned}
C[4,1] &= \infty \\
C[4,2] &= \infty \\
C[4,3] &= U_a(4) + U_b(3) + P_{ab}(4,3) \\
C[4,4] &= U_a(4) + U_b(4) \\
C[4,5] &= U_a(4) + U_b(5) + P_{ab}(4,5)
\end{aligned}
$$

$$
\begin{aligned}
C[5,1] &= \infty \\
C[5,2] &= \infty \\
C[5,3] &= \infty \\
C[5,4] &= U_a(5) + U_b(4) + P_{ab}(5,4) \\
C[5,5] &= U_a(5) + U_b(5)
\end{aligned}
$$

This graph construction only assigns finite costs to solutions where the neighboring pixels are the same or differ by one value, and hence ensures smooth solutions.

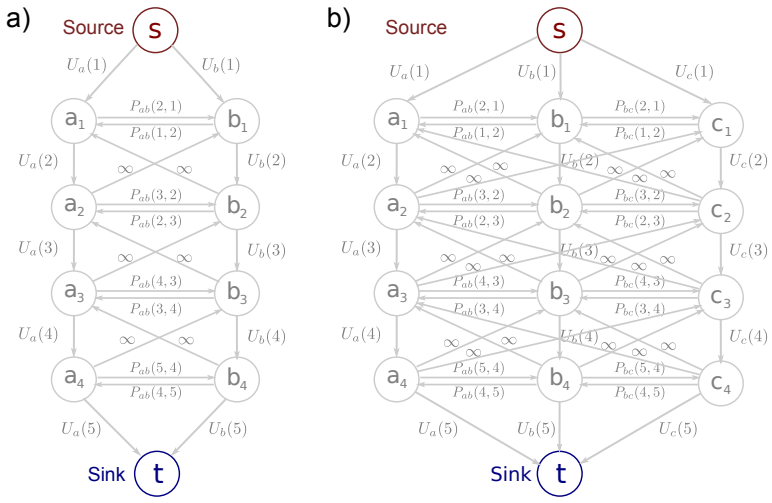

Figure 12.2 Alternative multi-label graph constructions. Each of these these graphs has extra *constraint links* with infinite weight. These have the effect of giving an infinite cost to a subset of the possible solutions.

Problem 12.6 Which of the possible minimum cuts of the graph in figure 12.2b have a finite cost?

Problem 12.7 Confirm that the costs of the cuts in figure 12.14b from the book are as claimed by explicitly performing the summation over the relevant terms C_{ij}.

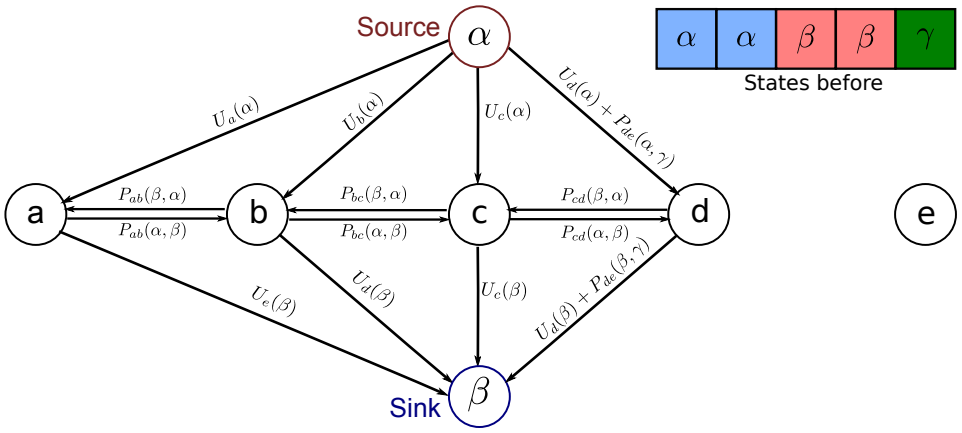

Figure 12.3 Graph construction for problem 12.9

Problem 12.8 Show that the Potts model (figure 12.17c) is not submodular by providing a counter-example to the required criterion:

$$P_{ab}(\beta, \gamma) + P_{ab}(\alpha, \delta) - P_{ab}(\beta, \delta) - P_{ab}(\alpha, \gamma) \geq 0.$$

Problem 12.9 An alternative to the alpha-expansion algorithm is the alpha-beta swap. Here, a multi-label MRF with non-convex potentials is optimized by repeatedly choosing pairs of labels α, β and performing a binary graph cut that allows them to swap in such a way that the overall cost function decreases. Devise a graph structure that can be used to perform this operation. Hint: consider separate cases for neighboring labels α, α, β, β, β, γ, α, γ and γ, γ where γ is a label that is neither α or β.

Answer

The graph construction is shown in figure 12.3. Nodes which are neither α or β do not need to be included in the graph. The pairwise term that encodes the cost for one of the nodes in the graph being next to a node with a value that is neither α or β can be included in the links connecting this node to the source and sink. Notice that this graph construction does not require the triangle inequality to be true.

Chapter 13

Image processing and feature extraction

Problem 13.1 Consider an 8-bit image in which the pixel values are evenly distributed in the range 0to 127, with no pixels taking a value of 128 or larger. Draw the cumulative histogram for this image. What will the histogram of pixel intensities look like after applying histogram equalization?

Problem 13.2 Consider a continuous image $f[i,j]$ and a continuous filter $f[m,n]$ In the continuous domain, the operation $f \otimes p$ of convolving an image with the filter is defined as:

$$f \otimes p = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p[i-m, j-n]f[m,n]\, dmdn.$$

Now consider two filters f and g. Prove that convolving the image first with f and then with g has the same effect as convolving f with g and then convolving the image with the result. In other words:

$$g \otimes (f \otimes p) = (g \otimes f) \otimes p.$$

Does this result extend to discrete images?

Answer

Let's define images a and b as

$$\begin{aligned} a &= f \otimes p = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p[i-m, j-n]f[m,n]\, dmdn \\ b &= g \otimes f = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f[i-m, j-n]g[m,n]\, dmdn. \end{aligned}$$

Now, we use the same formula again to compute the left hand side

$$\begin{aligned} g \otimes (f \otimes p) &= g \otimes a \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a[k-i, l-j]g[i,j]\, didj \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p[k-i-m, l-j-n]f[m,n]g[i,j]\, didjdmdn \end{aligned}$$

and the right hand side

$$\begin{aligned} (g \otimes f) \otimes p &= b \otimes p \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p[k-m', l-n']b[m',n']\, dm'dn' \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p[k-m', l-n']f[m'-i, n'-j]g[i,j]\, didjdm'dn' \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p[k-i-m, l-j-n]f[m,n]g[i,j]\, didjdmdn \end{aligned}$$

where we have made the changes of variables $m' = m+i$ and $n' = n+j$ between the last two lines. We notice that both expressions are the same as required.

This extends to infinite discrete images, but we can get different results for finite sized images, depending on how we deal with the boundaries.

Problem 13.3 Describe the series of operations that would be required to compute the Haar-like filters in figures 13.6a-d from an integral image. How many points from the integral image are needed to compute each?

Problem 13.4 Consider a blurring filter where each pixel in an image is replaced by a weighted average of local intensity values, but the the weights decrease if these intensity values differ markedly from the central pixel. What effect would this *bilateral filter* have when applied to an image?

Problem 13.5 Define a 3×3 filter that is specialized to detecting luminance changes at a 45^o angle and gives a positive response where the image intensity increases from the bottom left to the bottom right of the image.

Problem 13.6 Define a 3×3 filter that responds to the second derivative in the horizontal direction, but is invariant to the gradient and absolute intensity in the horizontal direction and invariant to all changes in the vertical direction.

Problem 13.7 Why are most local binary patterns in a natural image typically uniform or near-uniform?

Problem 13.8 Give one example of a 2D data set where the mixtures of Gaussians model will succeed in clustering the data, but the K-means algorithm will fail.

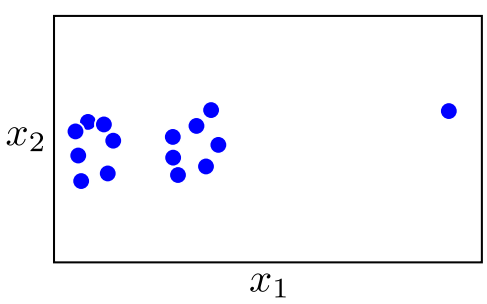

Figure 13.1 Clustering with the K-means algorithm in the presence of outliers (problem 13.9). This data set contains two clusters and a single outlier (the point on the right hand side). The outlier causes problems for the K-means algorithm when $K = 2$ clusters are used due to the implicit assumption that the clusters can be modeled as normal distributions with spherical covariance.

Problem 13.9 Consider the data in figure 13.1. What do you expect to happen if we run the K-means algorithm with two clusters on this data set? Suggest a way to resolve this problem.

Answer

The outlying point will drag the mean of one of the clusters into the large blank space and the result will be that the central data cluster will wind up being shared by the two components. One way to handle this would be to use a mixture of two t-distributions which are robust to outliers.

Problem 13.10 An alternative approach to clustering the data would be to find modes (peaks) in the density of the points. This potentially has the advantage of also automatically selecting the number of clusters. Propose an algorithm to find these modes.

Answer

One simple approach is the 'mean-shift' algorithm. We determine a starting position and place a spheroid (circle in 2D) centered at this position. The radius should be such that it includes a number of data points, but not all of them. Then we move the center of the spheroid to a new position that is the average of the points that fall within the radius. We iterate this process. The result is that the spheroid gradually moves towards regions of greater density and will eventually stop moving when it reaches a mode. We can repeat this from many starting positions to find all of the modes.

Chapter 14

The pinhole camera

Problem 14.1 A pinhole camera has a sensor that is $1cm \times 1cm$ and a horizontal field of view of 60^o. What is the distance between the optical center and the sensor? The same camera has a resolution of 100 pixels in the horizontal direction and 200 pixels in the vertical direction (i.e., the pixels are not square). What are the focal length parameters f_x and f_y from the intrinsic matrix?

Problem 14.2 We can use the pinhole camera model to understand a famous movie effect. *Dolly zoom* was first used in Alfred Hitchcock's *Vertigo*. As the protagonist looks down a stairwell, it appears to deform (figure 14.1) in a strange way. The background seems to move away from the camera, while the foreground remains at a constant position.

In terms of the camera model, two things occur simultaneously during the dolly zoom sequence: the camera moves along the w-axis, and the focal distance of the camera changes. The distance moved and the change of focal length are carefully chosen so that objects in a pre-defined plane remain at the same position. However, objects out of this plane move relative to one another (figures 14.1d-e).

I want to capture two pictures of a scene at either end of a 'dolly zoom'. Before the zoom, the camera is at $w = 0$, the distance between the optical center and the image plane is $1cm$, and the image plane is $1cm \times 1cm$. After the zoom, the camera is at $w = 100cm$. I want the plane at $w = 500cm$ to be stable after the camera movement. What should the new distance between the optical center and the image plane be?

Problem 14.3 Figure 14.2 shows two different camera models: the orthographic and weak perspective cameras. For each camera, devise the relationship between the homogeneous world points and homogeneous image points. You may assume that the world coordinate system and the camera coordinate system coincide, so there is no need to introduce the extrinsic matrix.

Problem 14.4 Find the point where the homogeneous lines $\tilde{\mathbf{l}}_1$ and $\tilde{\mathbf{l}}_2$ join where:

1. $\tilde{\mathbf{l}}_1 = [3, 1, 1]$, and $\tilde{\mathbf{l}}_2 = [-1, 0, 1]$
2. $\tilde{\mathbf{l}}_1 = [1, 0, 1]$, and $\tilde{\mathbf{l}}_2 = [3, 0, 1]$

Hint: the (3×1) homogeneous point vector $\tilde{\mathbf{x}}$ must satisfy both $\tilde{\mathbf{l}}_1^T\tilde{\mathbf{x}} = 0$ and $\tilde{\mathbf{l}}_2^T\tilde{\mathbf{x}} = 0$. In other words it should be orthogonal to both $\tilde{\mathbf{l}}_1$ and $\tilde{\mathbf{l}}_2$.

Answer

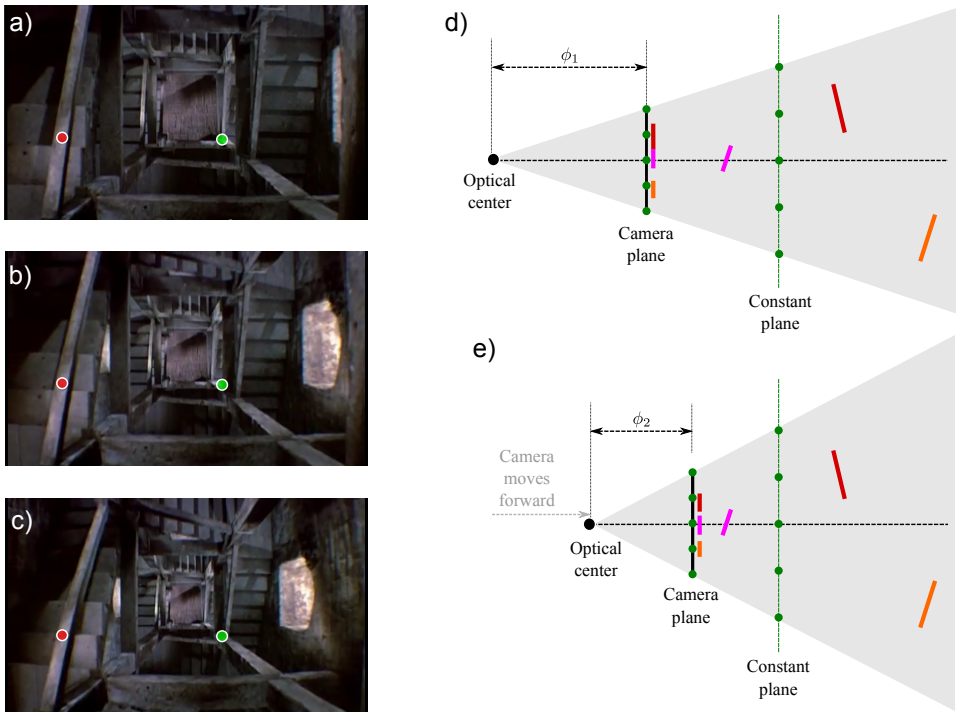

Figure 14.1 Dolly Zoom. (a-c) Three frames from 'Vertigo' sequence in which the stairwell appears to distort. Nearby objects remain in roughly the same place whereas object further away systematically move through the sequence. To see this, consider the red and green circles that are at the same (x, y) position in each frame. The red circle remains on the near bannister, but the green circle is on the floor of the stairwell in the first image but halfway up the stairs in the last image. d) To understand this effect consider a camera viewing a scene which consists of several green points at the same depth and some other surfaces (colored lines). e) We move the camera along the w-axis but simultaneously change the focal length so that the green points are imaged at the same position. Under these changes, objects in the plane of the green points are static, but other parts of the scene move and may even occlude one another.

1. To find a vector that is orthogonal to both $\tilde{\mathbf{l}}_1$ and $\tilde{\mathbf{l}}_2$ we take the cross product

$$\tilde{\mathbf{x}} = \tilde{\mathbf{l}}_1 \times \tilde{\mathbf{l}}_2 = \begin{vmatrix} i & j & k \\ 3 & 1 & 1 \\ -1 & 0 & 1 \end{vmatrix} = \begin{bmatrix} -1 \\ 4 \\ -1 \end{bmatrix}$$

2. By the same process, we get

$$\tilde{\mathbf{x}} = \tilde{\mathbf{l}}_1 \times \tilde{\mathbf{l}}_2 = \begin{vmatrix} i & j & k \\ 1 & 0 & 1 \\ 3 & 0 & 1 \end{vmatrix} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$$

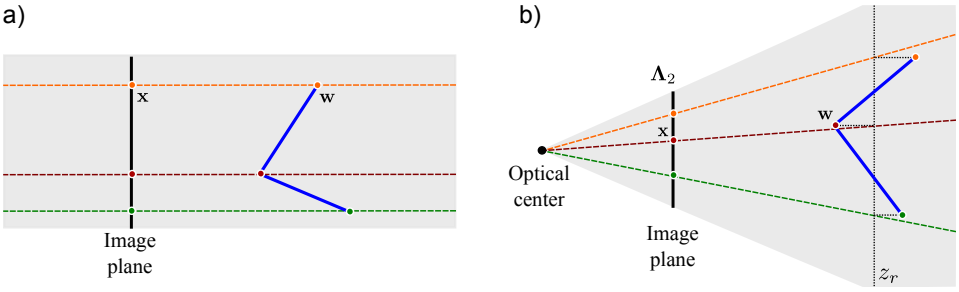

Figure 14.2 Alternative camera models. a) Orthographic camera. Rays are parallel and orthogonal to image plane. b) Weak perspective model. Points are projected orthogonally onto a reference plane at distance z_r from the camera and then pass to the image plane by perspective projection.

Notice that these two lines were parallel, so the point where the converge is at infinity (last component of homogeneous representation is zero).

Problem 14.5 Find the line joining the homogeneous points $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ where

$$\tilde{\mathbf{x}}_1 = [2, 2, 1]^T, \tilde{\mathbf{x}}_2 = [-2, -2, 1]^T.$$

Problem 14.6 A conic $\mathbf{C}$ is a geometric structure that can represent ellipses and circles in the 2D image. The condition for a point to lie on a conic is given by

$$\begin{bmatrix} x & y & 1 \end{bmatrix} \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = 0,$$

or

$$\tilde{\mathbf{x}}^T \mathbf{C} \tilde{\mathbf{x}} = 0.$$

Describe an algorithm to estimate the parameters a, b, c, d, e, f given several points $x_1, x_2, \ldots, x_n$ that are known to lie on the conic. What is the minimum number of points that your algorithm requires to be successful?

This has the form $\mathbf{A}\mathbf{w} = 0$ of a minimum direction problem (see Appendix C of book) and can be solved using the singular value decomposition. The six unknowns are ambiguous up to scale, so we need at least 5 points to find a unique solution.

Problem 14.7 Devise a method to find the intrinsic matrix of a projector using a camera and known calibration object.

Problem 14.8 What is the minimum number of binary striped light patterns of the type illustrated in figure 14.13 required to estimate the camera-projector correspondences for a projector image of size $H \times W$?

Problem 14.9 There is a potential problem with the shape from silhouette algorithm as described; the point that we have found on the surface of the object may be occluded by another part of the object with respect to the nearest camera. Consequently, when we copy the color, we will get the wrong value. Propose a method to circumvent this problem.

Problem 14.10 In the augmented reality application (figure **??**), the realism might be enhanced if the object had a shadow. Propose an algorithm that could establish whether a point on the desktop (assumed planar) is shadowed by the object with respect to a point light source at a known position.

Chapter 15

Models for transformations

Problem 15.1 The 2D point $\mathbf{x}_2$ is created by a rotating point $\mathbf{x}_1$ using the rotation matrix $\boldsymbol{\Omega}_1$ and then translating it by the translation vector $\boldsymbol{\tau}_1$ so that

$$\mathbf{x}_2 = \boldsymbol{\Omega}_1\mathbf{x}_1 + \boldsymbol{\tau}_1.$$

Find the parameters $\boldsymbol{\Omega}_2$ and $\boldsymbol{\tau}_2$ of the inverse transformation

$$\mathbf{x}_1 = \boldsymbol{\Omega}_2\mathbf{x}_2 + \boldsymbol{\tau}_2$$

in terms of the original parameters $\boldsymbol{\Omega}_1$ and $\boldsymbol{\tau}_1$.

Problem 15.2 A 2D line can be as expressed as $ax + by + c = 0$ or in homogeneous terms

$$\mathbf{l}\tilde{\mathbf{x}} = 0,$$

where $\mathbf{l} = [a, b, c]$. If points are transformed so that

$$\tilde{\mathbf{x}}' = \mathbf{T}\tilde{\mathbf{x}},$$

what is the equation of the transformed line?

Problem 15.3 Using your solution from problem 15.2, develop a linear algorithm for estimating a homography based on a number of matched lines between the two images (i.e., the analogue of the DLT algorithm for matched lines).

Answer

Using the result of the previous question, the relation between the two sets of lines is

$$\mathbf{l}\boldsymbol{\Phi}^{-1} = \mathbf{l}'$$

or, re-arranging by post-multiplying by $\boldsymbol{\Phi}$ and taking the transpose we have

$$\mathbf{l}^T = \boldsymbol{\Phi}^T\mathbf{l}^T.$$

We can now use the relation

$$\mathbf{l}^T \times \boldsymbol{\Phi}^T\mathbf{l}^T = \mathbf{0}.$$

to create two independent equations for the elements of $\boldsymbol{\Phi}$ and proceed as in the DLT algorithm.

Problem 15.4 A conic (see problem 14.6) is defined by

$$\tilde{\mathbf{x}}^T\mathbf{C}\tilde{\mathbf{x}} = 0,$$

where $\mathbf{C}$ is a 3×3 matrix. If the points in the image undergo the transformation

$$\tilde{\mathbf{x}}' = \mathbf{T}\tilde{\mathbf{x}},$$

then what is the equation of the transformed conic?

Problem 15.5 All of the 2D transformations in this chapter (Euclidean, similarity, affine, projective) have 3D equivalents. For each class write out the 4×4 matrix that describes the 3D transformation in homogeneous coordinates. How many independent parameters does each model have?

Problem 15.6 Devise an algorithm to estimate a 3D affine transformation based on two sets of matching 3D points. What is the minimum number of points required to get a unique estimate of the parameters of this model?

Problem 15.7 A 1D affine transformation acts on 1D points x as $x' = ax + b$. Show that the ratio of two distances is *invariant* to a 1D affine transformation so that

$$I = \frac{x_1 - x_2}{x_2 - x_3} = \frac{x_1' - x_2'}{x_2' - x_3'}.$$

Problem 15.8 A 1D projective transform acts on 1D points x as $x' = (ax + b)/(cx + d)$. Show that the *cross-ratio* of distances is *invariant* to a 1D projective transformation so that

$$I = \frac{(x_3 - x_1)(x_4 - x_2)}{(x_3 - x_2)(x_4 - x_1)} = \frac{(x_3' - x_1')(x_4' - x_2')}{(x_3' - x_2')(x_4' - x_1')}.$$

Problem 15.9 Show that equation 15.36 follows from equation 15.35.

Answer

We start with the premise

$$\tilde{\mathbf{x}} \times \boldsymbol{\Phi}\tilde{\mathbf{w}} = \mathbf{0}$$

We can write $\boldsymbol{\Phi}\tilde{\mathbf{w}}$ as a vector

$$\boldsymbol{\Phi}\tilde{\mathbf{w}} = \begin{bmatrix} \boldsymbol{\phi}_{\bullet 1}\tilde{\mathbf{w}} \\ \boldsymbol{\phi}_{\bullet 2}\tilde{\mathbf{w}} \\ \boldsymbol{\phi}_{\bullet 3}\tilde{\mathbf{w}} \end{bmatrix}$$

where $\boldsymbol{\phi}_{\bullet,k}$ represents the k^{th} row of the matrix $\boldsymbol{\Phi}$.

Now we take the cross product to give

$$\begin{aligned}
\tilde{\mathbf{x}} \times \boldsymbol{\Phi}\tilde{\mathbf{w}} &= \begin{vmatrix} i & j & k \\ x & y & 1 \\ \boldsymbol{\phi}_{\bullet 1}\tilde{\mathbf{w}} & \boldsymbol{\phi}_{\bullet 2}\tilde{\mathbf{w}} & \boldsymbol{\phi}_{\bullet 3}\tilde{\mathbf{w}} \end{vmatrix} \\
&= \begin{bmatrix} y\boldsymbol{\phi}_{\bullet 3}\tilde{\mathbf{w}} - \boldsymbol{\phi}_{\bullet 2}\tilde{\mathbf{w}} \\ \boldsymbol{\phi}_{\bullet 1}\tilde{\mathbf{w}} - x\boldsymbol{\phi}_{\bullet 3}\tilde{\mathbf{w}} \\ x\boldsymbol{\phi}_{\bullet 2}\tilde{\mathbf{w}} - y\boldsymbol{\phi}_{\bullet 1}\tilde{\mathbf{w}} \end{bmatrix} \\
&= \begin{bmatrix} y(\phi_{31}u + \phi_{32}v + \phi_{33}) - (\phi_{21}u + \phi_{22}v + \phi_{23}) \\ (\phi_{11}u + \phi_{12}v + \phi_{13}) - x(\phi_{31}u + \phi_{32}v + \phi_{33}) \\ x(\phi_{21}u + \phi_{22}v + \phi_{23}) - y(\phi_{11}u + \phi_{12}v + \phi_{13}) \end{bmatrix} = \mathbf{0}.
\end{aligned}$$

as required.

Problem 15.10 A camera with intrinsic matrix $\boldsymbol{\lambda}$ and extrinsic parameters $\boldsymbol{\Omega} = \mathbf{I}, \boldsymbol{\tau} = \mathbf{0}$ takes an image and then rotates to a new position $\boldsymbol{\Omega} = \boldsymbol{\Omega}_1, \boldsymbol{\tau} = \mathbf{0}$ and takes a second image. Show that the homography relating these two images is given by

$$\boldsymbol{\Phi} = \boldsymbol{\Lambda}\boldsymbol{\Omega}_1\boldsymbol{\Lambda}^{-1}.$$

Problem 15.11 Consider two images of the same scene taken from a camera that rotates, but does not translate between taking the images. What is the minimum number of point matches required to recover a 3D rotation between two images taken using a camera where the intrinsic matrix is known?

Problem 15.12 Consider the problem of computing a homography from point matches that include outliers. If 50% of the initial matches are correct, how many iterations of the RANSAC algorithm would we expect to have to run in order to have a 95% chance of computing the correct homography?

Problem 15.13 A different approach to fitting transformations in the presence of outliers is to model the uncertainty as a mixture of two Gaussians. The first Gaussian models the image noise, and the second Gaussian, which has a very large variance, accounts for the outliers. For example, for the affine transformation we would have

$$Pr(\mathbf{x}|\mathbf{w}) = \lambda \text{Norm}_{\mathbf{x}}\left[\mathbf{aff}[\mathbf{w}, \boldsymbol{\Phi}, \boldsymbol{\tau}], \sigma^2\mathbf{I}\right] + (1-\lambda)\text{Norm}_{\mathbf{x}}\left[\mathbf{aff}[\mathbf{w}, \boldsymbol{\Phi}, \boldsymbol{\tau}], \sigma_0^2\mathbf{I}\right] +$$

where λ is the probability of being an inlier, σ^2 is the image noise and σ_0^2 is the large variance that accounts for the outliers. Sketch an approach to learning the parameters $\sigma^2, \boldsymbol{\Phi}, \boldsymbol{\tau}$, and λ of this model. You may assume that σ_0^2 is fixed. Identify a possible weakness of this model.

Problem 15.14 In the description of how to compute the panorama (section **??**), it is suggested that we take each pixel in the central image and transform it into the other images and then copy the color. What is wrong with the alternate strategy of taking each pixel from the other images and transforming them into the central image?

Chapter 16

Multiple cameras

Problem 16.1 Sketch the pattern of epipolar lines on the images in figure 16.2a.

Problem 16.2 Show that the cross product relation can be written in terms of a matrix multiplication so that

$$\mathbf{a} \times \mathbf{b} = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

Problem 16.3 Consider figure 16.1. Write the direction of the three 3D vectors $\mathbf{O}_1\mathbf{O}_2$, $\mathbf{O}_1\mathbf{w}$, and $\mathbf{O}_2\mathbf{w}$ in terms of the observed image positions $\mathbf{x}_1$,$\mathbf{x}_2$ and the rotation $\boldsymbol{\Omega}$ and translation $\boldsymbol{\tau}$ between the cameras. The scale of the vectors is unimportant.

The three vectors that you have found must be coplanar. The criterion for three 3D vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ being coplanar can be written as $\mathbf{a}.(\mathbf{b} \times \mathbf{c}) = 0$. Use this criterion to derive the essential matrix.

Answer

From the perspective of camera 2, the vector $O_1\mathbf{w} = \kappa_1\Omega\mathbf{x}_1$, the vector $O_1O_2 = \kappa_2\tau$ and the vector $O_2\mathbf{w} = \mathbf{x}_2$. Combining these using the coplanarity condition we get

$$\mathbf{x}_2^T(\tau \times \boldsymbol{\Omega}\mathbf{x}_1) = 0$$

as required.

Problem 16.4 A clueless computer vision professor writes:

"The essential matrix is a 3×3 matrix that relates image coordinates between two images of the same scene. It contains 8 independent degrees of freedom (it is ambiguous up to scale). It has rank 2. If we know the intrinsic matrices of the two cameras, we can use the essential matrix to recover the rotation and translation between the cameras exactly."

Edit this statement to make it factually correct.

Problem 16.5 The essential matrix relates points in two cameras so that

$$\mathbf{x}_2^T\mathbf{E}\mathbf{x}_1 = 0$$

is given by

$$\mathbf{E} = \begin{bmatrix} 0 & 0 & 10 \\ 0 & 0 & 0 \\ -10 & 0 & 0 \end{bmatrix}.$$

What is the epipolar line in image 2 corresponding to the point $x_1 = [1, -1, 1]$? What is the epipolar line in image 2 corresponding to the points $x_1 = [-5, -2, 1]$? Determine the position of the epipole in image 2. What can you say about the motion of the cameras?

Problem 16.6 Show that we can retrieve the essential matrix by multiplying together the expressions from the decomposition (equations 16.19) as $\mathbf{E} = \boldsymbol{\tau}_\times\boldsymbol{\Omega}$.

Problem 16.7 Derive the fundamental matrix relation:

$$\tilde{\mathbf{x}}_2^T\boldsymbol{\Lambda}_2^{-T}\mathbf{E}\boldsymbol{\Lambda}_1^{-1}\tilde{\mathbf{x}}_1 = 0.$$

Problem 16.8 I intend to compute the fundamental matrix using the eight-point algorithm. Unfortunately, my data set is polluted by 30% outliers. How many iterations of the RANSAC algorithm will I need to run to have a 99% probability of success (i.e., computing the fundamental matrix from eight inliers at least once)? How many iterations will I need if I use an algorithm based on seven points?

Problem 16.9 We are given the fundamental matrix $\mathbf{F}_{13}$ relating images 1 and 3 and the fundamental matrix $\mathbf{F}_{23}$ relating images 2 and 3. I am now given corresponding points $\mathbf{x}_1$ and $\mathbf{x}_2$ in images 1 and 2, respectively. Derive a formula for the position of the corresponding point in image 3.

Problem 16.10 Tomasi-Kanade factorization. In the orthographic camera (figure 14.2), the projection process can be described as

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} \phi_x & \gamma & \delta_x \\ 0 & \phi_y & \delta_y \end{bmatrix} \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \tau_x \\ \omega_{21} & \omega_{22} & \omega_{23} & \tau_y \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} + \begin{bmatrix} \tau_x' \\ \tau_y' \end{bmatrix}, \end{aligned}$$

or in matrix form

$$\mathbf{x} = \boldsymbol{\Pi}\mathbf{w} + \boldsymbol{\tau}'.$$

Now consider a data matrix $\mathbf{X}$ containing the positions $\{\mathbf{x}_{ij}\}_{i,j=1}^{IJ}$ of J points as seen in I images so that

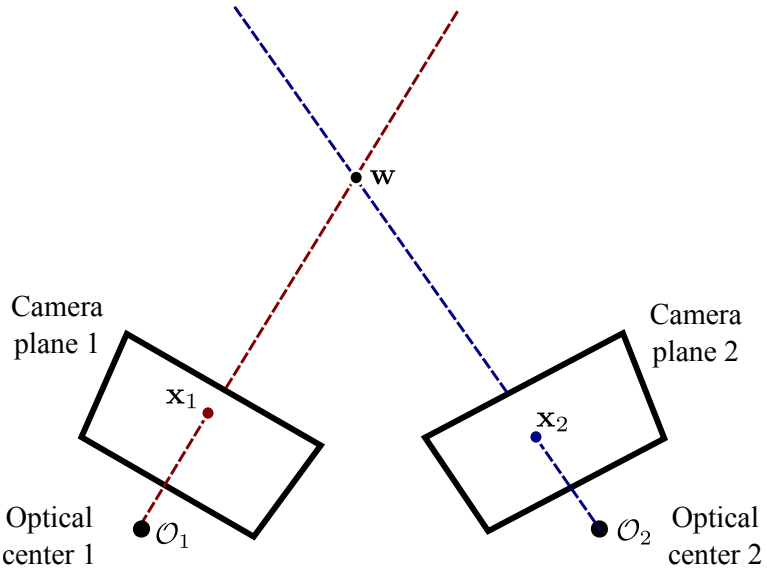

Figure 16.1 Figure for problem 16.3.

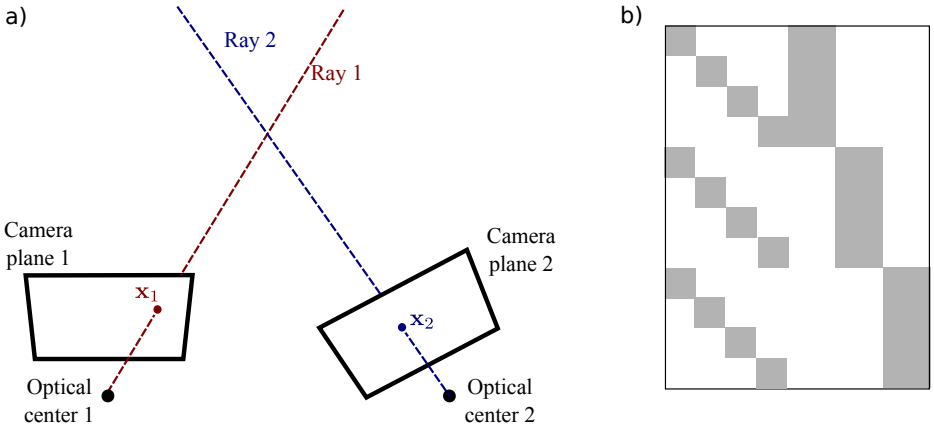

Figure 16.2 a) Figure for problem 16.1. c) Figure for problem 16.11. Gray regions represent non-zero entries in this portrait Jacobian matrix.

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \ldots & \mathbf{x}_{iJ} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \ldots & \mathbf{x}_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{I1} & \mathbf{x}_{I2} & \ldots & \mathbf{x}_{IJ}, \end{bmatrix}$$

and where $\mathbf{x}_{ij} = [x_{ij}, y_{ij}]^T$.

(i) Show that the matrix $\mathbf{X}$ can be written in the form

$$\mathbf{X} = \mathbf{PW} + \mathbf{T}$$

where $\mathbf{P}$ contains all of the I 3×2 projection matrices $\{\mathbf{\Pi}_i\}_{i=1}^{I}$, $\mathbf{W}$ contains all of the J 3D world positions $\{\mathbf{w}_j\}_{j=1}^{J}$ and $\mathbf{T}$ contains the translation vectors $\{\boldsymbol{\tau}_i'\}_{i=1}^{I}$.

(ii) Devise an algorithm to recover the matrices $\mathbf{P}, \mathbf{W}$ and $\mathbf{T}$ from the measurements $\mathbf{X}$. Is your solution unique?

Answer

i) We can define the matrices

$$\begin{aligned}
\mathbf{P} &= \begin{bmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_2 \\ \vdots \\ \boldsymbol{\pi}_I \end{bmatrix} \\
\mathbf{W} &= \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \ldots & \mathbf{w}_J \end{bmatrix} \\
\mathbf{T} &= \begin{bmatrix} \boldsymbol{\tau}_1 & \boldsymbol{\tau}_1 & \ldots & \boldsymbol{\tau}_1 \\ \boldsymbol{\tau}_2 & \boldsymbol{\tau}_2 \ldots \boldsymbol{\tau}_2 & & \\ \vdots & \vdots & \ldots & \vdots \\ \boldsymbol{\tau}_I & \boldsymbol{\tau}_I & \vdots & \boldsymbol{\tau}_i \end{bmatrix}
\end{aligned}$$

and then the condition $\mathbf{X} = \mathbf{PW} + \mathbf{T}$ will be fulfilled.

ii) A simple algorithm is to re-arrange the equation to get

$$\mathbf{X} - \mathbf{T} = \mathbf{PW}$$

and then observe that $\mathbf{PW}$ is the product of a $2I \times 3$ matrix and a $3 \times J$ matrix. To find the best approximation to these matrices, we can compute the svd

$$[\mathbf{U}, \mathbf{L}, \mathbf{V}] = \mathbf{svd}[\mathbf{X} - \mathbf{T}]$$

and then set all but the first 3 values of $\mathbf{L}$ to zero. We can now choose $\mathbf{P} = \mathbf{UL}$ and $\mathbf{W} = \mathbf{V}^T$.

Problem 16.11 Consider a Jacobian that has a structure of non-zero entries as shown in figure 16.2b. Draw an equivalent image that shows the structure of the non-zero entries in the matrix $\mathbf{J}^T\mathbf{J}$. Describe how you would use the Schur complement relation to invert this matrix efficiently.

Chapter 17

Models for shape

Problem 17.1 A conic is defined as the set of points where

$$\begin{bmatrix} x & y & 1 \end{bmatrix} \begin{bmatrix} \alpha & \beta & \gamma \\ \beta & \delta & \epsilon \\ \gamma & \epsilon & \zeta \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = 0,$$

or

$$\tilde{\mathbf{x}}^T \mathbf{C} \tilde{\mathbf{x}} = 0.$$

Use MATLAB to draw the 2D function $\tilde{\mathbf{x}}^T \mathbf{C} \tilde{\mathbf{x}}$ and identify the set of positions where this function is zero for the following matrices:

$$\mathbf{C}_1 = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix} \qquad \mathbf{C}_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & -2 \end{bmatrix} \qquad \mathbf{C}_3 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Problem 17.2 Devise an efficient algorithm to compute the distance transform. The algorithm should take a binary image and return at each pixel the city block distance to the nearest non-zero element of the original image. The city block distance d between pixels (x_1, y_1) and pixel (x_2, y_2) is defined as

$$d = |x_1 - x_2| + |y_1 - y_2|.$$

Problem 17.3 Consider a prior that is based in the curvature term:

$$\text{curve}[\mathbf{w}, n] = -(\mathbf{w}_{n-1} - 2\mathbf{w}_n + \mathbf{w}_{n+1})^T (\mathbf{w}_{n-1} - 2\mathbf{w}_n + \mathbf{w}_{n+1}).$$

If landmark point $\mathbf{w}_1 = [100, 100]$ and landmark point $\mathbf{w}_3$ is at position $\mathbf{w}_3 = [200, 300]$, what position $\mathbf{w}_2$ will minimize the function Curve$[\mathbf{w}, 2]$?

Problem 17.4 If the snake as described in section 17.2 is initialized in an empty image, how would you expect it to evolve during the fitting procedure?

Problem 17.5 The spacing element of the snake prior 17.5 encourages all of the control points of the snake to be the equidistant. An alternative approach is to give the snake a tendency to shrink (so that it collapses around objects). Write out an alternative expression for the spacing term that accomplishes this goal.

Problem 17.6 Devise a method to find the 'best' weight vector $\mathbf{h}$ given a new vector $\mathbf{w}$ and the parameters $\{\boldsymbol{\mu}, \boldsymbol{\Phi}, \sigma^2\}$ of the PPCA model (see figure **??**).

Problem 17.7 Show that if the singular value decomposition of a matrix $\mathbf{W}$ can be written as $\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{V}^T$, then it follows that

$$\begin{aligned}\mathbf{W}\mathbf{W}^T &= \mathbf{U}\mathbf{L}^2\mathbf{U}^T \\ \mathbf{W}^T\mathbf{W} &= \mathbf{V}\mathbf{L}^2\mathbf{V}^T.\end{aligned}$$

Answer

$$\begin{aligned}\mathbf{W}\mathbf{W}^T &= \mathbf{U}\mathbf{L}\mathbf{V}^T(\mathbf{U}\mathbf{L}\mathbf{V}^T)^T = \mathbf{U}\mathbf{L}\mathbf{V}^T\mathbf{V}\mathbf{L}\mathbf{U}^2 = \mathbf{U}\mathbf{L}^2\mathbf{U}^T \\ \mathbf{W}^T\mathbf{W} &= (\mathbf{U}\mathbf{L}\mathbf{V}^T)^T\mathbf{U}\mathbf{L}\mathbf{V}^T = \mathbf{V}\mathbf{L}\mathbf{U}^T\mathbf{U}\mathbf{L}\mathbf{V}^2 = \mathbf{V}\mathbf{L}^2\mathbf{V}^T.\end{aligned}$$

Problem 17.8 Devise a method to learn the PPCA model using the EM algorithm, giving details of both the E- and M-steps. Are you guaranteed to get the same answer as the method based on the SVD?

Problem 17.9 Show that the maximum a posteriori solution for the hidden weight variable $\mathbf{h}$ is as given in equation 17.32.

Problem 17.10 You are given a set of 100 male faces and 100 female faces. By hand you mark 50 landmark points on each image. Describe how to use this data to develop a generative approach to gender classification based on shape alone. Describe both the training process and how you would infer the gender for a new face that does not contain landmark points.

Problem 17.11 Imagine that we have learned a point distribution for the shape of the human face. Now we see a new face where everything below the nose is occluded by a scarf. How could you exploit the model to estimate both the positions of the landmark points in the top half of the face and the landmark points in the (missing) bottom half of the face?

Problem 17.12 An alternative approach to building a nonlinear model of shape is to use a mixture model. Describe an approach to training a statistical shape model based on the mixture of probabilistic principal component analyzers. How would you fit this model to a new image?

Problem 17.13 One way to warp one image to another is to implement a piecewise affine warp. Assume that we have a number of points in image 1 and their corresponding points in image 2. We first triangulate each set of points in the same way. We now represent the position $\mathbf{x}_1$ in image 1 as a weighted sum of the three vertices of the triangle $\mathbf{a}_1, \mathbf{b}_1, \mathbf{c}_1$ that it lies in so that

$$\mathbf{x}_1 = \alpha\mathbf{a}_1 + \beta\mathbf{b}_1 + \gamma\mathbf{c}_1,$$

where the weights are constrained to be positive with $\alpha + \beta + \gamma = 1$. These weights are known as *barycentric coordinates*.

To find the position in the second image, we then compute the position relative to the three vertices $\mathbf{a}_2, \mathbf{b}_2, \mathbf{c}_2$ of the warped triangle so that

$$\mathbf{x}_2 = \alpha\mathbf{a}_2 + \beta\mathbf{b}_2 + \gamma\mathbf{c}_2.$$

How can we compute the weights α, β, γ? Devise a method to warp the whole image in this manner.

Problem 17.14 Consider an ellipsoid in 3D space that is represented by the quadric

$$\tilde{\mathbf{w}}^T \begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^T & c \end{bmatrix} \tilde{\mathbf{w}} = 0,$$

where $\mathbf{A}$ is a 3×3 matrix, $\mathbf{b}$ is a 3×1 vector, and c is a scalar.

For a normalized camera we can write the world point $\tilde{\mathbf{w}}$ in terms of the image point $\tilde{\mathbf{x}}$ as $\tilde{\mathbf{w}} = [\tilde{\mathbf{x}}^T, s]^T$ where s is a scaling factor that determines the distance along the ray $\tilde{\mathbf{x}}$.

(i) Combine these conditions to produce a criterion that must be true for an image point $\tilde{\mathbf{x}}$ to lie within the projection of the conic.

(ii) The edge of the image of the conic is the locus of points for which there is a single solution for the distance s. Outside the conic there is no real solution for s and inside it there are two possible solutions corresponding to the front and back face of the quadric. Use this intuition to derive an expression for the conic in terms of $\mathbf{A}, \mathbf{b}$ and $\mathbf{c}$. If the camera has intrinsic matrix $\mathbf{\Lambda}$, what would the new expression for the conic be?

Answer

i) Combining these equations together, we have

$$cs^2 + 2\mathbf{b}^T\tilde{\mathbf{x}}s + \tilde{\mathbf{x}}^T\mathbf{A}\tilde{\mathbf{x}} = 0$$

ii) This is a quadratic in s which has a unique solution when the discriminant is zero, so that

$$\tilde{\mathbf{x}}^T(\mathbf{b}\mathbf{b}^T - c\mathbf{A})\mathbf{x} = 0$$

This is a conic with parameters $\mathbf{C} = \mathbf{b}\mathbf{b}^T - c\mathbf{A}$.

If the camera has intrinsic matrix Λ, then the points w now have the form $\tilde{\mathbf{w}} = [\Lambda^{-1}\tilde{\mathbf{x}}^T, s]^T$. Following this through the equations we get a conic with parameters $\mathbf{C} = \mathbf{\Lambda}^{-T}(\mathbf{b}\mathbf{b}^T - c\mathbf{A})\mathbf{\Lambda}^{-1}$.

Chapter 18

Models for style and identity

Problem 18.1 Prove that the posterior distribution over the hidden variable in the subspace identity model is as given in equation 18.9.

Problem 18.2 Show that the M-step updates for the subspace identity model are as given in equation 18.11.

Problem 18.3 Develop a closed form solution for learning the parameters $\{\boldsymbol{\mu}, \boldsymbol{\Phi}, \sigma^2\}$ of a subspace identity model where the noise is spherical:

$$Pr(\mathbf{x}_{ij}) = \text{Norm}_{\mathbf{x}_{ij}}[\boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \sigma^2\mathbf{I}].$$

Hint: Assume you have exactly $J = 2$ examples of each of the I training images and base your solution on probabilistic PCA.

Problem 18.4 In a face clustering problem, how many possible models of the data are there with 2,3,4,10, and 100 faces?

Problem 18.5 An alternative approach to face verification using the identity subspace model is to compute the probability of the observed data $\mathbf{x}_1$ and $\mathbf{x}_2$ under the models:

$$\begin{aligned} Pr(\mathbf{x}_1, \mathbf{x}_2|w = 0) &= Pr(\mathbf{x}_1)Pr(\mathbf{x}_2) \\ Pr(\mathbf{x}_1, \mathbf{x}_2|w = 1) &= Pr(\mathbf{x}_1)Pr(\mathbf{x}_2|\mathbf{x}_1). \end{aligned}$$

Write down expressions for the marginal probability terms $Pr(\mathbf{x}_1)$, $Pr(\mathbf{x}_2)$ and the conditional probability $Pr(\mathbf{x}_2|\mathbf{x}_1)$. How could you use these expressions to compute the posterior $Pr(w|\mathbf{x}_1, \mathbf{x}_2)$ over the world state?

Problem 18.6 Propose a version of the subspace identity model that is robust to outliers in the training data.

Problem 18.7 Moghaddam et al. (2000) took a different probabilistic approach to face verification. They took the difference $\mathbf{x}_\Delta = \mathbf{x}_2 - \mathbf{x}_1$ and modeled the likelihoods of this vector $Pr(\mathbf{x}_\Delta|w = 0)$ and $Pr(\mathbf{x}_\Delta|w = 1)$ when the two faces match or don't. Propose expressions for these likelihoods and discuss learning and inference in this model. Identify one possible disadvantage of this model.

Answer

The likelihoods of the difference image in each situation can be described as factor analyzers

$$Pr(\mathbf{x}_\delta|w) = \text{Norm}_{\mathbf{x}_\delta}[\boldsymbol{\mu}_w, \boldsymbol{\Phi}_w\boldsymbol{\Phi}_w^T + \sigma_w^2\mathbf{I}].$$

Inference can be achieved using Bayes' rule

$$Pr(\mathbf{w}|\mathbf{x}_\delta) = \frac{Pr(\mathbf{x}_\delta|w)Pr(w)}{Pr(\mathbf{x}_\delta)}$$

where $Pr(w)$ is a Bernoulli distribution that assigns a prior probability to each of the two world states.

Problem 18.8 Develop a model that combines the advantages of PLDA and the asymmetric bilinear model; it should be able to model the within-individual covariance with a subspace, but also be able to compare data between disparate styles. Discuss learning and inference in your model.

Problem 18.9 In the asymmetric bilinear model, how would you infer whether the style of two examples is the same or not, regardless of whether the images matched?

Chapter 19

Temporal models

Problem 19.1 Prove the Chapman-Kolmogorov relation:

$$\begin{aligned} Pr(\mathbf{w}_t|\mathbf{x}_{1\ldots t-1}) &= \int Pr(\mathbf{w}_t|\mathbf{w}_{t-1})Pr(\mathbf{w}_{t-1}|\mathbf{x}_{1\ldots t-1})\, d\mathbf{w}_{t-1} \\ &= \int \text{Norm}_{\mathbf{w}_t}[\boldsymbol{\mu}_p + \boldsymbol{\Psi}\mathbf{w}_{t-1}, \boldsymbol{\Sigma}_p]\text{Norm}_{\mathbf{w}_{t-1}}[\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}]\, d\mathbf{w}_{t-1} \\ &= \text{Norm}_{\mathbf{w}_t}[\boldsymbol{\mu}_p + \boldsymbol{\Psi}\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_p + \boldsymbol{\Psi}\boldsymbol{\Sigma}_{t-1}\boldsymbol{\Psi}^T] \end{aligned}$$

Answer

It is possible to solve this integral directly, using the relations in chapter 5. However, taking on trust that the result will be a normal distribution, we can directly compute the mean and covariance. We first write the expression as a generative equation

$$w_t = \boldsymbol{\mu}_p + \boldsymbol{\Psi}\mathbf{w}_{t-1} + \epsilon_p$$

where ϵ_p is noise with mean $\boldsymbol{\mu}_{t-1}$ and covariance $\boldsymbol{\Sigma}_{t-1}$.
The mean is given by

$$\begin{aligned} E[\mathbf{w}_t] &= E[\boldsymbol{\mu}_p + \boldsymbol{\Psi}\mathbf{w}_{t-1} + \epsilon_p] \\ &= \boldsymbol{\mu}_p + \boldsymbol{\Psi}\mathbf{w}_{t-1} + 0 \\ &= \boldsymbol{\mu}_p + \boldsymbol{\Psi}\mathbf{w}_{t-1} \end{aligned}$$

as required.
The covariance is given by

$$\begin{aligned} E[(w_t - \overline{\mathbf{w}}_t)(w_t - \overline{\mathbf{w}}_t)^T] &= E[\boldsymbol{\Psi}(\mathbf{w}_{t-1} - \boldsymbol{\mu}_{t-1}) + \epsilon_p)(\boldsymbol{\Psi}(\mathbf{w}_{t-1} - \boldsymbol{\mu}_{t-1}) + \epsilon_p)^T] && (19.1) \\ &= E[\boldsymbol{\Psi}[(\mathbf{w}_{t-1} - \boldsymbol{\mu}_{t-1})(\mathbf{w}_{t-1} - \boldsymbol{\mu}_{t-1})^T]\boldsymbol{\Psi}^T + E[\epsilon_p\epsilon_p^T] + \boldsymbol{\Psi}^T E[(\mathbf{w}_{t-1} - \boldsymbol{\mu}_{t-1})\epsilon_p^T] && (19.2) \\ &= \boldsymbol{\Psi}\boldsymbol{\Sigma}_{t-1}\boldsymbol{\Psi}^T + \boldsymbol{\Sigma}_p && (19.3) \end{aligned}$$

as required.

Problem 19.2 Derive the measurement incorporation step for the Kalman filter. In other words, show that

$$\begin{aligned} Pr(\mathbf{w}_t|\mathbf{x}_{1\ldots t}) &= \frac{Pr(\mathbf{x}_t|\mathbf{w}_t)Pr(\mathbf{w}_t|\mathbf{w}_{1\ldots t-1},\mathbf{x}_{1\ldots t})}{Pr(\mathbf{x}_{1\ldots t})} \\ &= \frac{\text{Norm}_{\mathbf{x}_t}[\boldsymbol{\mu}_m+\boldsymbol{\Phi}\mathbf{w}_t,\boldsymbol{\Sigma}_m]\text{Norm}_{\mathbf{w}_t}[\boldsymbol{\mu}_+,\boldsymbol{\Sigma}_+]}{Pr(\mathbf{x}_{1\ldots t})} \\ &= \text{Norm}_{\mathbf{w}_t}\left[\left(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\Phi}+\boldsymbol{\Sigma}_+^{-1}\right)^{-1}\left(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_m)+\boldsymbol{\Sigma}_+^{-1}\boldsymbol{\mu}_+\right),\right. \\ &\qquad\left.\left(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\Phi}+\boldsymbol{\Sigma}_+^{-1}\right)^{-1}\right]. \end{aligned}$$

Answer

We begin by applying the change in variables relation (section 5.7) which gives us

$$\begin{aligned} Pr(\mathbf{w}_t|\mathbf{x}_{1\ldots t}) &\propto \text{Norm}_{\mathbf{x}_t}[\boldsymbol{\mu}_m+\boldsymbol{\Phi}\mathbf{w}_t,\boldsymbol{\Sigma}_m]\text{Norm}_{\mathbf{w}_t}[\boldsymbol{\mu}_+,\boldsymbol{\Sigma}_+] \\ &= \text{Norm}_{\mathbf{w}_t}[(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_m,(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\Phi})^{-1})]\text{Norm}_{\mathbf{w}_t}[\boldsymbol{\mu}_+,\boldsymbol{\Sigma}_+] \end{aligned}$$

Now we apply the product of two normal distributions rule (section 5.6) which gives us

$$\begin{aligned} Pr(\mathbf{w}_t|\mathbf{x}_{1\ldots t}) &\propto \text{Norm}_{\mathbf{w}_t}[(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_m,(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\Phi})^{-1})]\text{Norm}_{\mathbf{w}_t}[\boldsymbol{\mu}_+,\boldsymbol{\Sigma}_+] \\ &= \text{Norm}_{\mathbf{w}_t}[(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\Phi}+\boldsymbol{\Sigma}_+)^{-1}(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})+\boldsymbol{\Sigma}_+^{-1}\boldsymbol{\mu}_+),(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\Phi}+\boldsymbol{\Sigma}_+)^{-1}] \end{aligned}$$

as required. Note that the constant must be one to make the posterior a valid probability distribution.

Problem 19.3 Consider a Kalman filter type system where the prior based on the previous time step is a mixture of K Gaussians

$$Pr(\mathbf{w}_t|\mathbf{x}_{1\ldots t-1}) = \sum_{k=1}^{K}\lambda_k\text{Norm}_{\mathbf{w}_t}[\boldsymbol{\mu}_{+k},\boldsymbol{\Sigma}_{+k}].$$

What will happen in the subsequent measurement incorporation step? What will happen in the next time update step?

Problem 19.4 Consider a model where there are two possible temporal update equations, represented by state transition matrices $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ and the system periodically switches from one regime to the other. Write a set of equations that describe this model and discuss a strategy for max-marginals inference.

Problem 19.5 In a Kalman filter model, discuss how you would compute the joint posterior distribution $Pr(\mathbf{w}_{1\ldots T}|\mathbf{x}_{1\ldots T})$ over all of the unknown world states. What form will this posterior distribution take? In the Kalman filter we choose to compute the marginal posteriors instead. Why is this?

Answer

The marginal posterior takes the form of a joint normal distribution between all of the variables. It can be solved for in closed form, by writing out the joint likelihood of all of the data conditioned on the hidden variables, and the joint prior distribution of all of the hidden variables and then using Bayes' rule. However, in practice, we usually have a large number of measurements so this is a very large system of equations and solving it is computationally intensive.

Problem 19.6 Apply the sum-product algorithm (section **??**) to the Kalman filter model and show that the result is equivalent to applying the Kalman filter recursions.

Problem 19.7 Prove the Kalman smoother recursions:

$$\begin{aligned}\boldsymbol{\mu}_{t|T} &= \boldsymbol{\mu}_t + \mathbf{C}_t(\boldsymbol{\mu}_{t+1|T} - \boldsymbol{\mu}_{+|t}) \\ \boldsymbol{\Sigma}_{t|T} &= \boldsymbol{\Sigma}_t + \mathbf{C}_t(\boldsymbol{\Sigma}_{t+1|T} - \boldsymbol{\Sigma}_{+|t})\mathbf{C}_t^T,\end{aligned}$$

where

$$\mathbf{C}_t = \boldsymbol{\Sigma}_{t|t}\boldsymbol{\Psi}^T\boldsymbol{\Sigma}_{+|t}^{-1}.$$

Hint: It may help to examine the proof of the forward-backward algorithm for HMMs (section 11.4.2).

Problem 19.8 Discuss how you would learn the parameters of the Kalman filter model given training sequences consisting of (i) both the known world state and the observed data, and (ii) just the observed data alone.

Problem 19.9 In the unscented Kalman filter we represented a Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ with a set of weights

$$\begin{aligned}\hat{\mathbf{w}}^{[0]} &= \boldsymbol{\mu}_{t-1} \\ \hat{\mathbf{w}}^{[j]} &= \boldsymbol{\mu}_{t-1} + \sqrt{\frac{D_{\mathbf{w}}}{1-a_0}}\boldsymbol{\Sigma}_{t-1}^{1/2}\mathbf{e}_j \qquad \text{for all } j \in \{1 \ldots D_{\mathbf{w}}\} \\ \hat{\mathbf{w}}^{[D_{\mathbf{w}}+j]} &= \boldsymbol{\mu}_{t-1} - \sqrt{\frac{D_{\mathbf{w}}}{1-a_0}}\boldsymbol{\Sigma}_{t-1}^{1/2}\mathbf{e}_j \qquad \text{for all } j \in \{1 \ldots D_{\mathbf{w}}\},\end{aligned}$$

with associated weights

$$a_j = \frac{1-a_0}{2D_{\mathbf{w}}}.$$

Show that the mean and covariance of these points are indeed $\boldsymbol{\mu}_{t-1}$ and $\boldsymbol{\Sigma}_{t-1}$ so that

$$\begin{aligned}\boldsymbol{\mu}_{t-1} &= \sum_{j=0}^{2D_{\mathbf{w}}} a_j\hat{\mathbf{w}}^{[j]} \\ \boldsymbol{\Sigma}_{t-1} &= \sum_{j=0}^{2D_{\mathbf{w}}} a_j(\hat{\mathbf{w}}^{[j]} - \boldsymbol{\mu}_{t-1})(\hat{\mathbf{w}}^{[j]} - \boldsymbol{\mu}_{t-1})^T.\end{aligned}$$

Problem 19.10 The extended Kalman filter requires the Jacobian matrix describing how small changes in the data create small changes in the measurements. Compute the Jacobian matrix for the measurement model for the pedestrian-tracking application (equation 19.50).

Chapter 20

Models for visual words

Problem 20.1 The bag of words method in this chapter uses a generative approach to model the frequencies of the visual words. Develop a discriminative approach that models the probability of the object class as a function of the word frequencies.

Problem 20.2 Prove the relations in equation 20.8, which show how to learn the latent Dirichlet allocation model in the case where we do know the part labels $\{p_{ij}\}_{i=1,j=1}^{I,J_i}$.

Answer

From equation 4.35 in the book, we have that the predictive density for a categorical model trained with data with frequencies $\{N_k\}_{k=1}^{K}$ is given by

$$Pr(x^* = k|\mathbf{x}_{1\ldots I}) = \frac{N_k + \alpha_k}{\sum_{j=1}^{6}(N_j + \alpha_j)}$$

where $\boldsymbol{\alpha}$ is hyperparameter of the Dirichlet prior.
Both of the terms in equation 20.8 take this form.

Problem 20.3 Show that the likelihood and prior terms in Latent Dirichlet Allocation are given by equations 20.10 and 20.11 respectively.

Answer

From problem 3.10 we have:

$$\prod_{i=1}^{I} \text{Cat}_{\mathbf{x}_i}[\lambda_{1\ldots K}] \cdot \text{Dir}_{\lambda_{1\ldots K}}[\alpha_{1\ldots K}] = \kappa \cdot \text{Dir}_{\lambda_{1\ldots K}}[\tilde{\alpha}_{1\ldots K}],$$

where

$$\begin{aligned} \tilde{\kappa} &= \frac{\Gamma[\sum_{j=1}^{K}\alpha_j]}{\Gamma[I+\sum_{j=1}^{K}\alpha_j]} \cdot \frac{\prod_{j=1}^{K}\Gamma[\alpha_j + N_j]}{\prod_{j=1}^{K}\Gamma[\alpha_j]} \\ \tilde{\alpha}_{1\ldots K} &= [\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K]. \end{aligned}$$

and N_k is the total number of times that the variable took the value k.

It follows that if we integrate over the unknown variables, the distribution integrates to one and we are left with the constant

$$\int \prod_{i=1}^{I} \text{Cat}_{\mathbf{x}_i}[\lambda_{1\ldots K}] \cdot \text{Dir}_{\lambda_{1\ldots K}}[\alpha_{1\ldots K}] = \kappa = \frac{\Gamma[\sum_{j=1}^{K}\alpha_j]}{\Gamma[I+\sum_{j=1}^{K}\alpha_j]} . \frac{\prod_{j=1}^{K}\Gamma[\alpha_j+N_j]}{\prod_{j=1}^{K}\Gamma[\alpha_j]}$$

It can now easily be seen that both equations 20.10 and 20.11 take this form.

Problem 20.4 Li and Perona (2005) developed an alternative model to the single author-topic model in which the hyperparameter $\boldsymbol{\alpha}$ was different for each value of the object label $\mathbf{w}$. Modify the graphical model for latent Dirichlet allocation to include this change.

Problem 20.5 Write out generative equations for the author-topic model in which multiple authors are allowed for each document. Draw the associated graphical model.

Problem 20.6 In real objects, we might expect visual words f that are adjacent to one another to take the same part label. How would you modify the author topic model to encourage nearby part labels to be the same. How would the Gibbs sampling procedure for drawing samples from the posterior probability over parts be affected?

Problem 20.7 Draw a graphical model for the scene model described in section 20.6.

Problem 20.8 All of the models in this chapter have dealt with classification; we wish to infer a discrete variable representing the state of the world based on discrete observed features $\{f_j\}$. Develop a generative model that can be used to infer a continuous variable based on discrete observed features (i.e., a regression model that uses visual words).