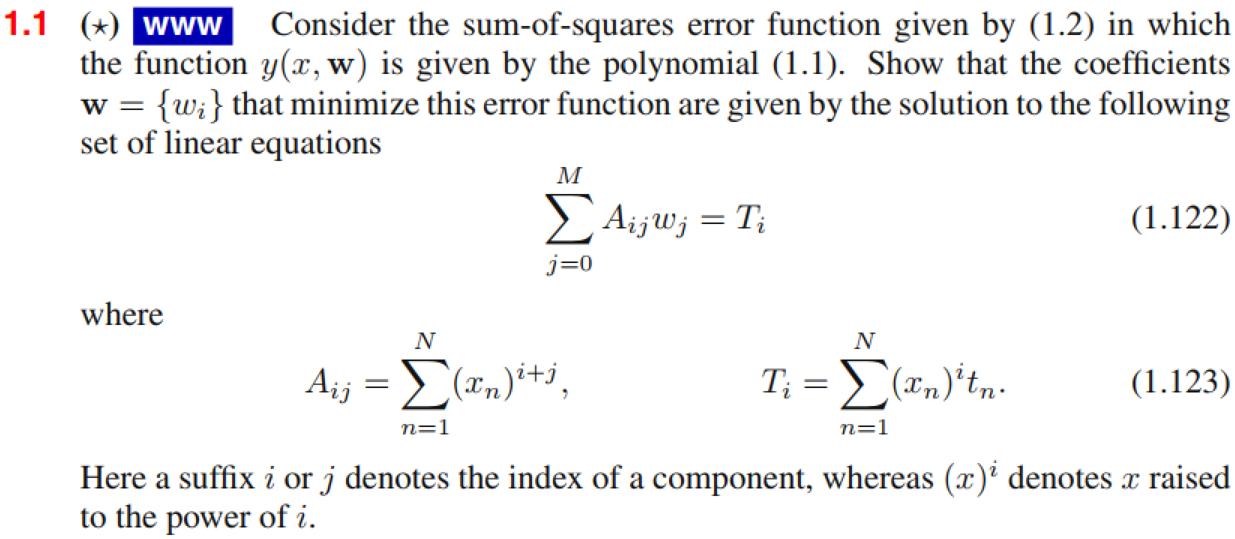
ECE1513
Tutorial 1:
Selected Exercises from Chapters 1 and 2

TA: Faye Pourghasem, fateme.pourghasem@mail.utoronto.ca

September 16, 2023

1 Example 1.1



1.1 ($\star$) **www** Consider the sum-of-squares error function given by (1.2) in which the function $y(x, \mathbf{w})$ is given by the polynomial (1.1). Show that the coefficients $\mathbf{w} = \{w_i\}$ that minimize this error function are given by the solution to the following set of linear equations

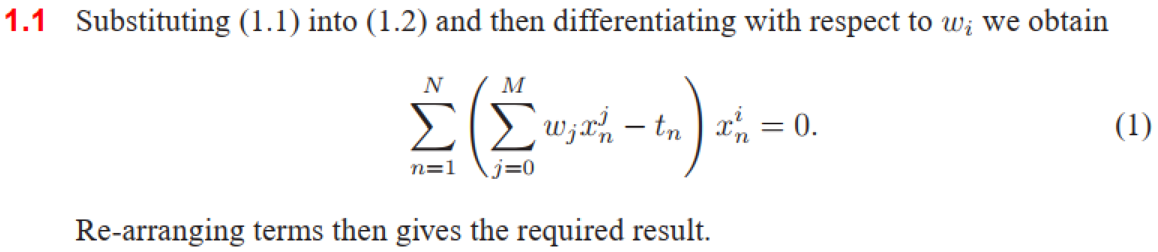
$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{1.122}$$

where

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}, \qquad T_i = \sum_{n=1}^{N} (x_n)^i t_n. \tag{1.123}$$

Here a suffix i or j denotes the index of a component, whereas $(x)^i$ denotes x raised to the power of i.

Solution

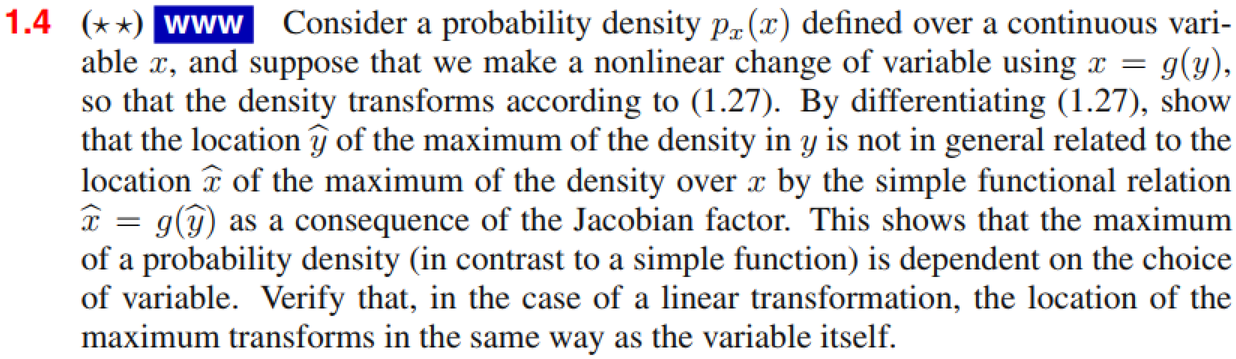


1.1 Substituting (1.1) into (1.2) and then differentiating with respect to w_i we obtain

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

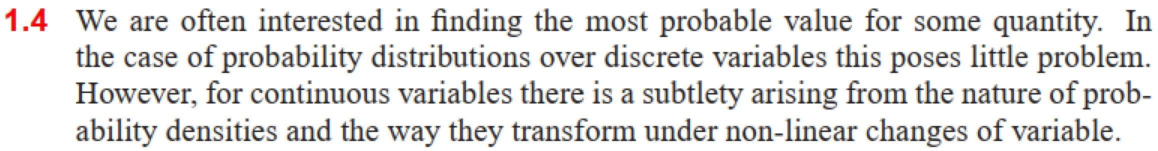
Re-arranging terms then gives the required result.

2 Example 1.4



1.4 ($\star\star$) **www** Consider a probability density $p_x(x)$ defined over a continuous variable x, and suppose that we make a nonlinear change of variable using $x = g(y)$, so that the density transforms according to (1.27). By differentiating (1.27), show that the location $\widehat{y}$ of the maximum of the density in y is not in general related to the location $\widehat{x}$ of the maximum of the density over x by the simple functional relation $\widehat{x} = g(\widehat{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent on the choice of variable. Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

Solution



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

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

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

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

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

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

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

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

Differentiating both sides with respect to y then gives

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

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

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

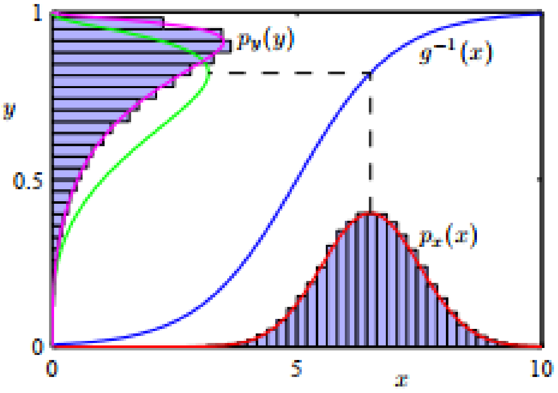
Now consider a non-linear change of variables from x to y given by

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

The inverse of this function is given by

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

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

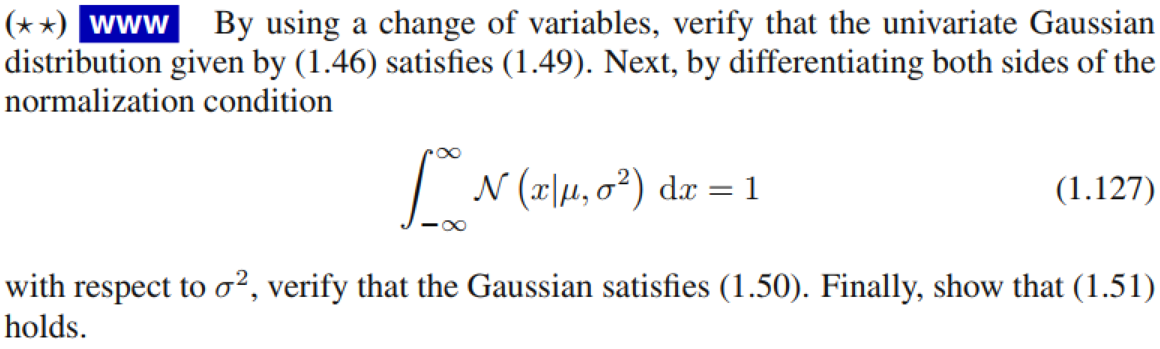


which is a *logistic sigmoid* function, and is shown in Figure 1 by the blue curve.

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

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

3 Example 1.8

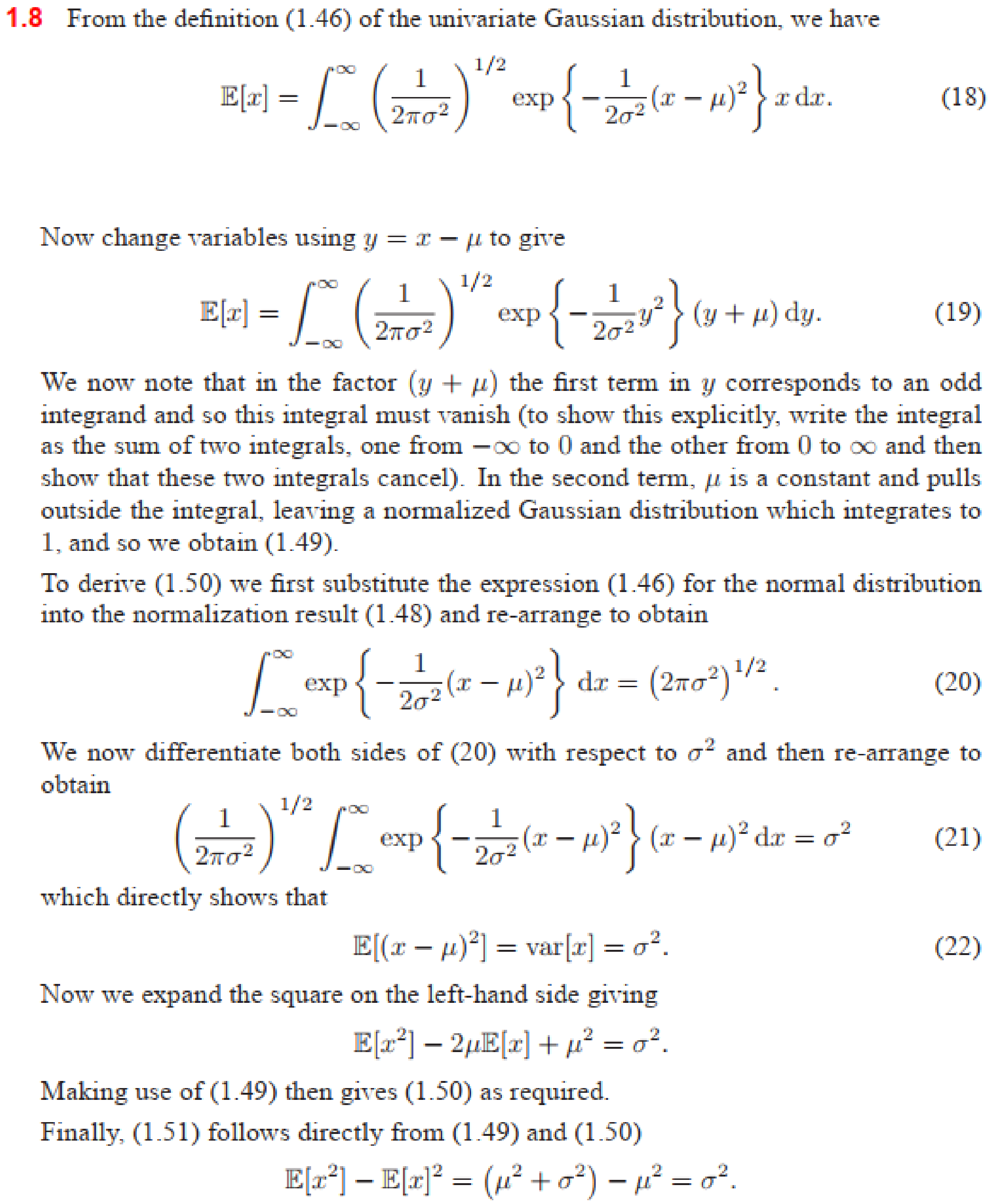


(⋆⋆) **www** By using a change of variables, verify that the univariate Gaussian distribution given by (1.46) satisfies (1.49). Next, by differentiating both sides of the normalization condition

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1 \tag{1.127}$$

with respect to σ^2, verify that the Gaussian satisfies (1.50). Finally, show that (1.51) holds.

Solution



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

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

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

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

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

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

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

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

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

which directly shows that

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

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

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

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

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

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

4 Example 1.20

1.20 $(\star\star)$ **www** In this exercise, we explore the behaviour of the Gaussian distribution in high-dimensional spaces. Consider a Gaussian distribution in D dimensions given by

$$p(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right). \tag{1.147}$$

We wish to find the density with respect to radius in polar coordinates in which the direction variables have been integrated out. To do this, show that the integral of the probability density over a thin shell of radius r and thickness ϵ, where $\epsilon \ll 1$, is given by $p(r)\epsilon$ where

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{1.148}$$

where S_D is the surface area of a unit sphere in D dimensions. Show that the function $p(r)$ has a single stationary point located, for large D, at $\widehat{r} \simeq \sqrt{D}\sigma$. By considering $p(\widehat{r}+\epsilon)$ where $\epsilon \ll \widehat{r}$, show that for large D,

$$p(\widehat{r}+\epsilon) = p(\widehat{r}) \exp\left(-\frac{3\epsilon^2}{2\sigma^2}\right) \tag{1.149}$$

which shows that $\widehat{r}$ is a maximum of the radial probability density and also that $p(r)$ decays exponentially away from its maximum at $\widehat{r}$ with length scale σ. We have already seen that $\sigma \ll \widehat{r}$ for large D, and so we see that most of the probability mass is concentrated in a thin shell at large radius. Finally, show that the probability density $p(\mathbf{x})$ is larger at the origin than at the radius $\widehat{r}$ by a factor of $\exp(D/2)$. We therefore see that most of the probability mass in a high-dimensional Gaussian distribution is located at a different radius from the region of high probability density. This property of distributions in spaces of high dimensionality will have important consequences when we consider Bayesian inference of model parameters in later chapters.

Solution

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

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

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

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

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

Next we note that

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

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

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

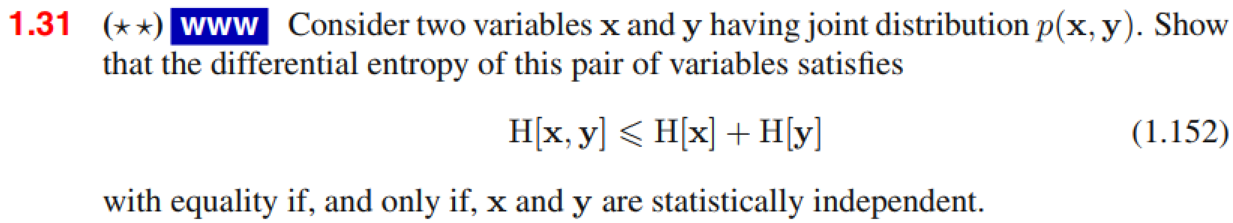
$$p(\mathbf{x} = \mathbf{0}) = \frac{1}{(2\pi\sigma^2)^{1/2}}$$

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

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

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

5 Example 1.31

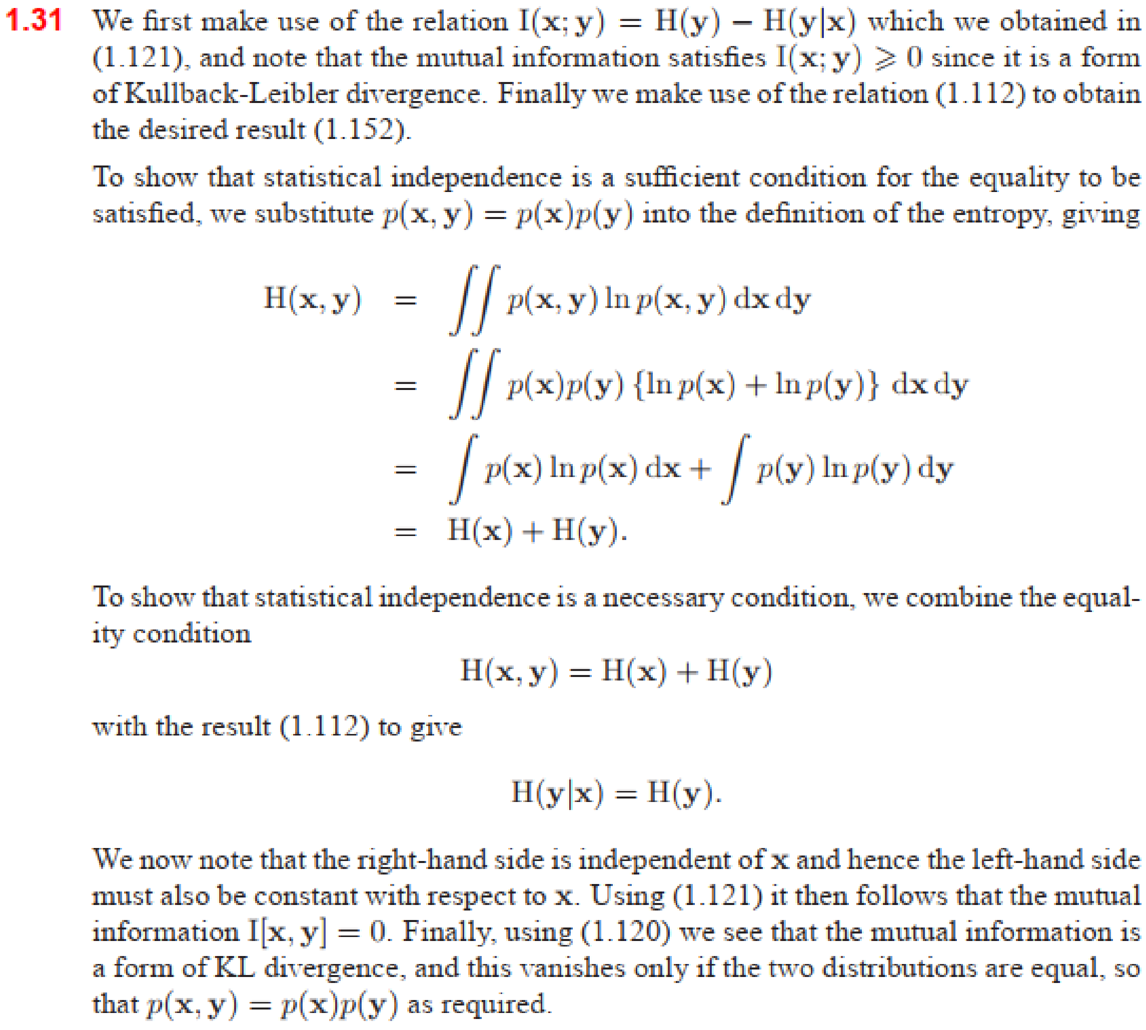


1.31 ($\star\star$) **www** Consider two variables $\mathbf{x}$ and $\mathbf{y}$ having joint distribution $p(\mathbf{x}, \mathbf{y})$. Show that the differential entropy of this pair of variables satisfies

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] \leqslant \mathrm{H}[\mathbf{x}] + \mathrm{H}[\mathbf{y}] \tag{1.152}$$

with equality if, and only if, $\mathbf{x}$ and $\mathbf{y}$ are statistically independent.

Solution



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

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

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

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

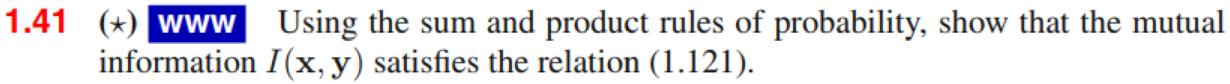
$$\mathrm{H}(\mathbf{x}, \mathbf{y}) = \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y})$$

with the result (1.112) to give

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

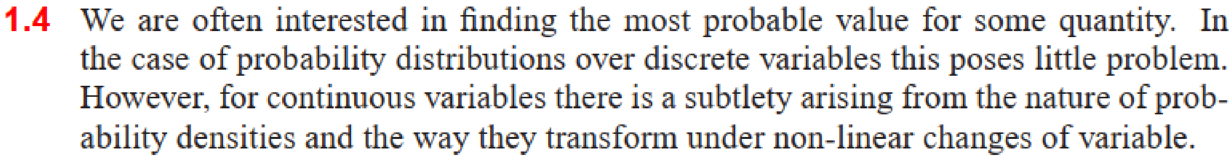
We now note that the right-hand side is independent of $\mathbf{x}$ and hence the left-hand side must also be constant with respect to $\mathbf{x}$. Using (1.121) it then follows that the mutual information $\mathrm{I}[\mathbf{x}, \mathbf{y}] = 0$. Finally, using (1.120) we see that the mutual information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

6 Example 1.41



1.41 $(\star)$ www Using the sum and product rules of probability, show that the mutual information $I(\mathbf{x}, \mathbf{y})$ satisfies the relation (1.121).

Solution



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

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

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

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

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

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

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

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

Differentiating both sides with respect to y then gives

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

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

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

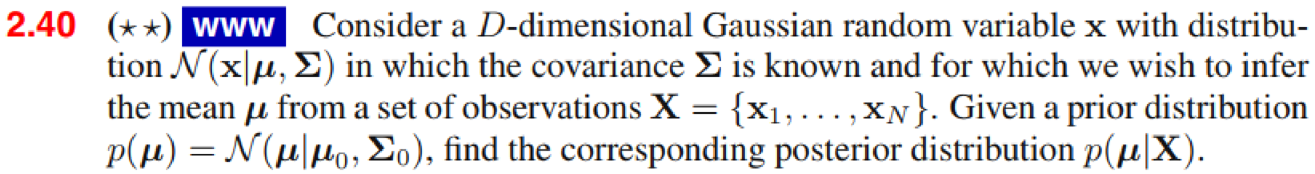
Now consider a non-linear change of variables from x to y given by

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

The inverse of this function is given by

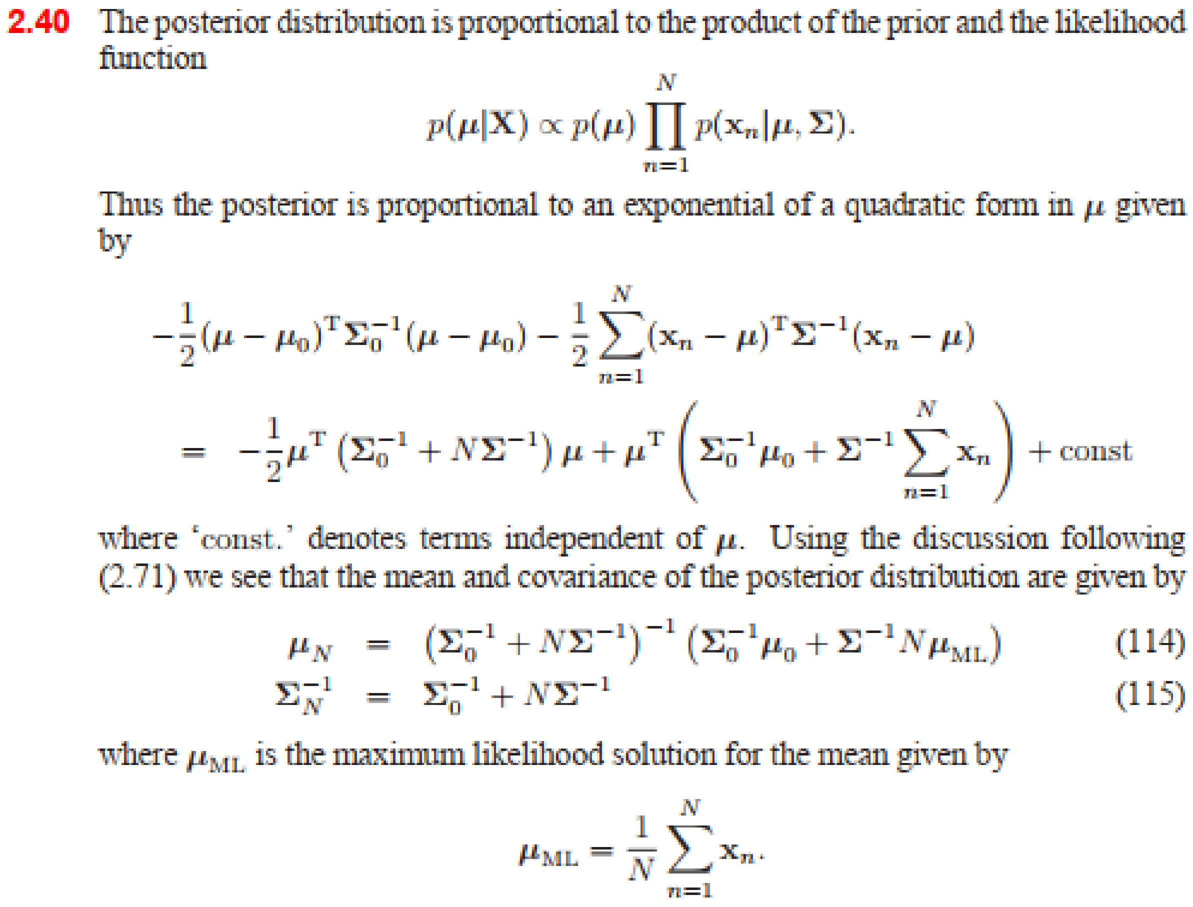
$$y = g^{-1}(x) = \frac{1}{1 + \exp(-x + 5)} \tag{11}$$

7 Example 2.4



2.40 $(\star\star)$ www Consider a D-dimensional Gaussian random variable $\mathbf{x}$ with distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ in which the covariance $\boldsymbol{\Sigma}$ is known and for which we wish to infer the mean $\boldsymbol{\mu}$ from a set of observations $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$. Given a prior distribution $p(\boldsymbol{\mu})=\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$, find the corresponding posterior distribution $p(\boldsymbol{\mu}|\mathbf{X})$.

Solution



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

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

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

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

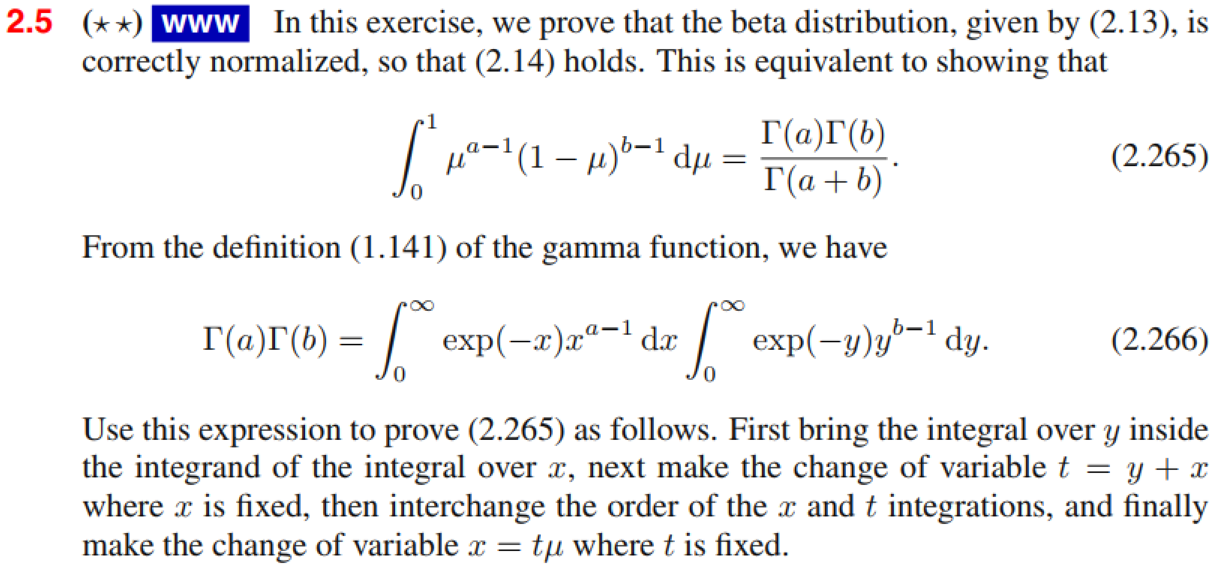
where 'const.' denotes terms independent of $\boldsymbol{\mu}$. Using the discussion following (2.71) we see that the mean and covariance of the posterior distribution are given by

$$\begin{aligned} \boldsymbol{\mu}_N &= \left(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\boldsymbol{\Sigma}^{-1}N\boldsymbol{\mu}_{\mathrm{ML}}\right) && (114) \\ \boldsymbol{\Sigma}_N^{-1} &= \boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1} && (115) \end{aligned}$$

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

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

8 Example 2.5



2.5 (⋆⋆) WWW In this exercise, we prove that the beta distribution, given by (2.13), is correctly normalized, so that (2.14) holds. This is equivalent to showing that

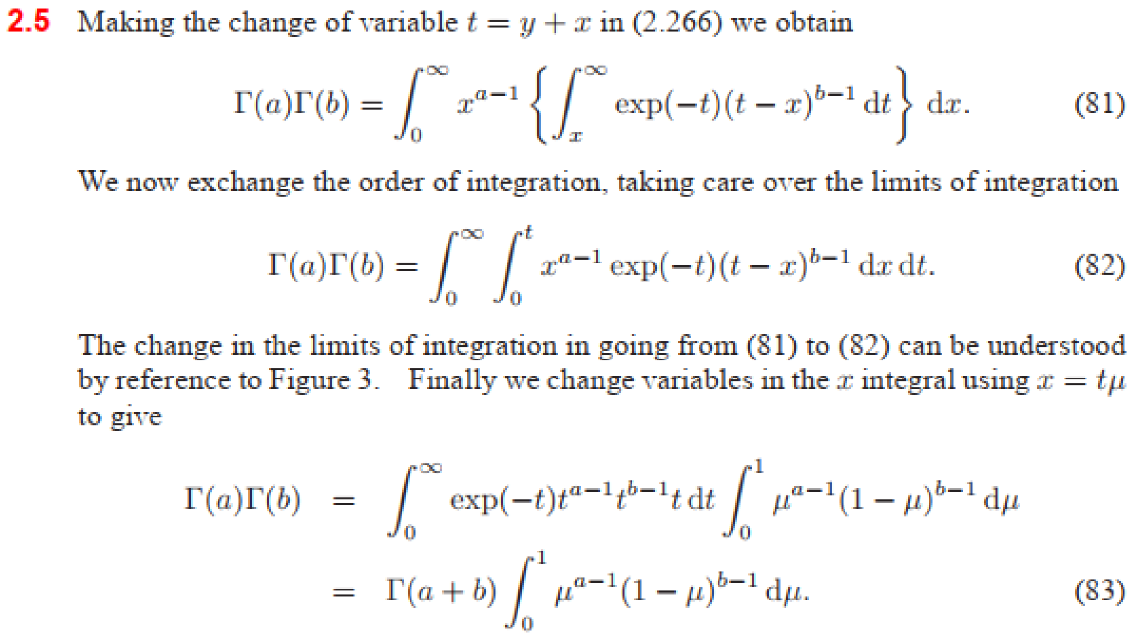
$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{2.265}$$

From the definition (1.141) of the gamma function, we have

$$\Gamma(a)\Gamma(b) = \int_0^\infty \exp(-x)x^{a-1}\,\mathrm{d}x \int_0^\infty \exp(-y)y^{b-1}\,\mathrm{d}y. \tag{2.266}$$

Use this expression to prove (2.265) as follows. First bring the integral over y inside the integrand of the integral over x, next make the change of variable $t = y + x$ where x is fixed, then interchange the order of the x and t integrations, and finally make the change of variable $x = t\mu$ where t is fixed.

Solution



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

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

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

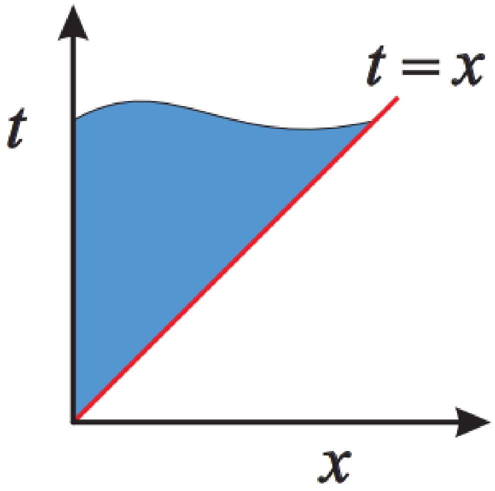
$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^t x^{a-1}\exp(-t)(t-x)^{b-1}\,\mathrm{d}x\,\mathrm{d}t. \tag{82}$$

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

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

Solutions 2.6–2.7 31

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



9 Example 2.14

2.14 $(\star\star)$ **www** This exercise demonstrates that the multivariate distribution with maximum entropy, for a given covariance, is a Gaussian. The entropy of a distribution $p(\mathbf{x})$ is given by

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{2.279}$$

We wish to maximize $\mathrm{H}[\mathbf{x}]$ over all distributions $p(\mathbf{x})$ subject to the constraints that $p(\mathbf{x})$ be normalized and that it have a specific mean and covariance, so that

$$\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1 \tag{2.280}$$

$$\int p(\mathbf{x})\mathbf{x}\,\mathrm{d}\mathbf{x} = \boldsymbol{\mu} \tag{2.281}$$

$$\int p(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\mathbf{x} = \boldsymbol{\Sigma}. \tag{2.282}$$

By performing a variational maximization of (2.279) and using Lagrange multipliers to enforce the constraints (2.280), (2.281), and (2.282), show that the maximum likelihood distribution is given by the Gaussian (2.43).

Solution

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

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

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

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

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

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

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

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

We now make the change of variable

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

The constraint (2.281) then becomes

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

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

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

and hence we have

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

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

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

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

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

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

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

10 Example 2.15

2.15 $(\star\star)$ Show that the entropy of the multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$\mathrm{H}[\mathbf{x}] = \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{D}{2}(1 + \ln(2\pi)) \qquad (2.283)$$

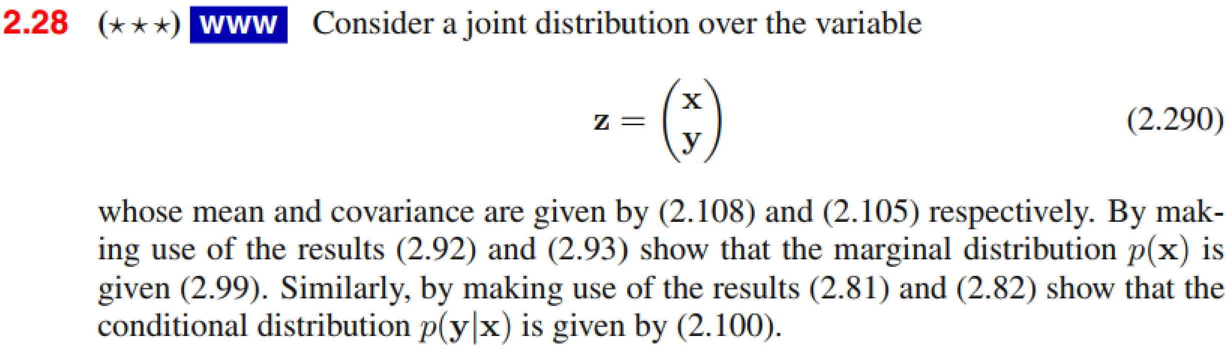
where D is the dimensionality of $\mathbf{x}$.

Solution

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

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

11 Example 2.28

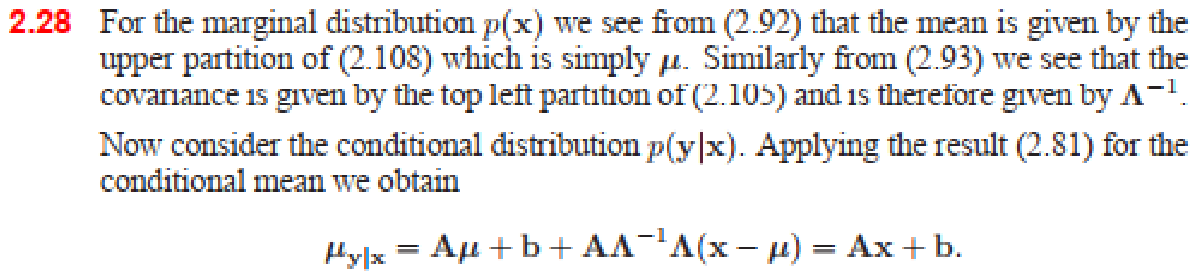


2.28 $(\star\star\star)$ www Consider a joint distribution over the variable

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \tag{2.290}$$

whose mean and covariance are given by (2.108) and (2.105) respectively. By making use of the results (2.92) and (2.93) show that the marginal distribution $p(\mathbf{x})$ is given (2.99). Similarly, by making use of the results (2.81) and (2.82) show that the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is given by (2.100).

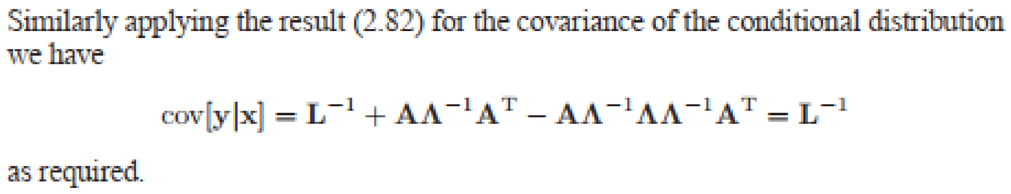
Solution



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

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

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

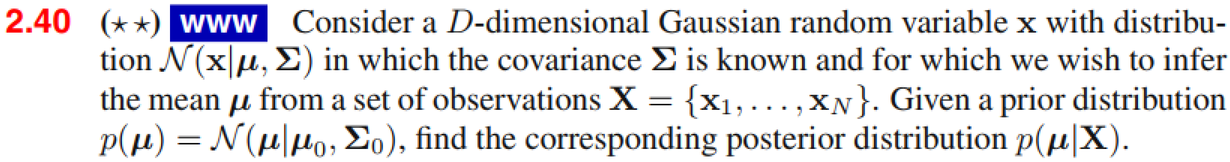


Similarly applying the result (2.82) for the covariance of the conditional distribution we have

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

as required.

12 Example 2.40



2.40 $(\star\star)$ www Consider a D-dimensional Gaussian random variable $\mathbf{x}$ with distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in which the covariance $\boldsymbol{\Sigma}$ is known and for which we wish to infer the mean $\boldsymbol{\mu}$ from a set of observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Given a prior distribution $p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, find the corresponding posterior distribution $p(\boldsymbol{\mu}|\mathbf{X})$.

Solution

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

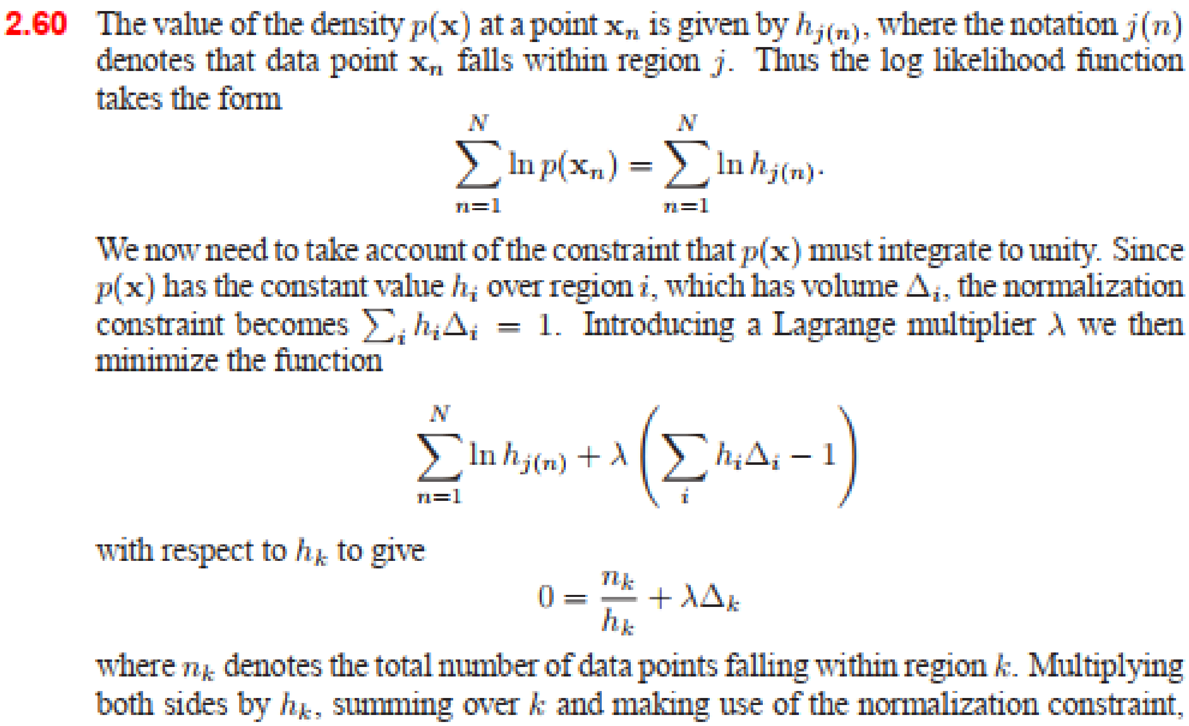
where $\boldsymbol{\mu}_{\mathrm{ML}}$ is the maximum likelihood solution for the mean given by

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

13 Example 2.6

2.60 $(\star\star)$ **www** Consider a histogram-like density model in which the space $\mathbf{x}$ is divided into fixed regions for which the density $p(\mathbf{x})$ takes the constant value h_i over the i^{th} region, and that the volume of region i is denoted Δ_i. Suppose we have a set of N observations of $\mathbf{x}$ such that n_i of these observations fall in region i. Using a Lagrange multiplier to enforce the normalization constraint on the density, derive an expression for the maximum likelihood estimator for the $\{h_i\}$.

Solution



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

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

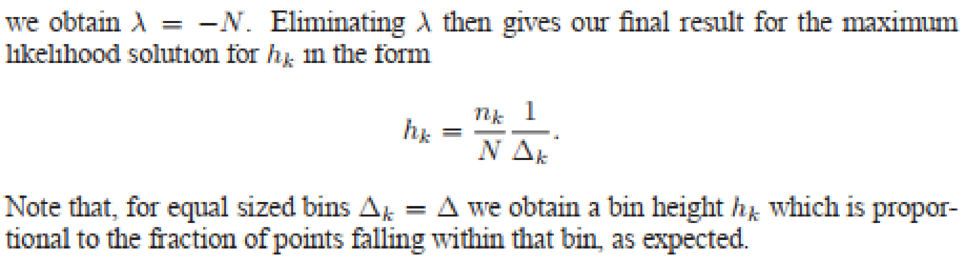
We now need to take account of the constraint that $p(\mathbf{x})$ must integrate to unity. Since $p(\mathbf{x})$ has the constant value h_i over region i, which has volume Δ_i, the normalization constraint becomes $\sum_i h_i \Delta_i = 1$. Introducing a Lagrange multiplier λ we then minimize the function

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

with respect to h_k to give

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

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



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

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

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