Chapter 5

The normal distribution

The most common representation for uncertainty in machine vision is the multivariate normal distribution. We devote this chapter to exploring its main properties, which will be used extensively throughout the rest of the book.

Recall from chapter 3 that the multivariate normal distribution has two parameters: the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The mean $\boldsymbol{\mu}$ is a $D\times 1$ vector that describes the position of the distribution. The covariance $\boldsymbol{\Sigma}$ is a symmetric $D\times D$ positive definite matrix (implying that $\mathbf{z}^T\boldsymbol{\Sigma}\mathbf{z}$ is positive for any real vector $\mathbf{z}$) and describes the shape of the distribution. The probability density function is

$$Pr(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-0.5(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{5.1}$$

or for short

$$Pr(\mathbf{x}) = \text{Norm}_{\mathbf{x}}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}\right]. \tag{5.2}$$

5.1 Types of covariance matrix

Covariance matrices in multivariate normals take three forms, termed *spherical*, *diagonal*, and *full* covariances. For the two dimensional (bivariate) case, these are

$$\boldsymbol{\Sigma}_{spher} = \begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix} \quad \boldsymbol{\Sigma}_{diag} = \begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix} \quad \boldsymbol{\Sigma}_{full} = \begin{bmatrix}\sigma_{11}^2 & \sigma_{12}^2\\ \sigma_{21}^2 & \sigma_{22}^2\end{bmatrix}. \tag{5.3}$$

The spherical covariance matrix is a positive multiple of the identity matrix and so has the same value on all of the diagonal elements and zeros elsewhere. In the diagonal covariance matrix, each value on the diagonal has a different positive value. The full covariance matrix can have non-zero elements everywhere although the matrix is still constrained to be symmetric and positive definite so for the 2D example, $\sigma_{12}^2 = \sigma_{21}^2$.

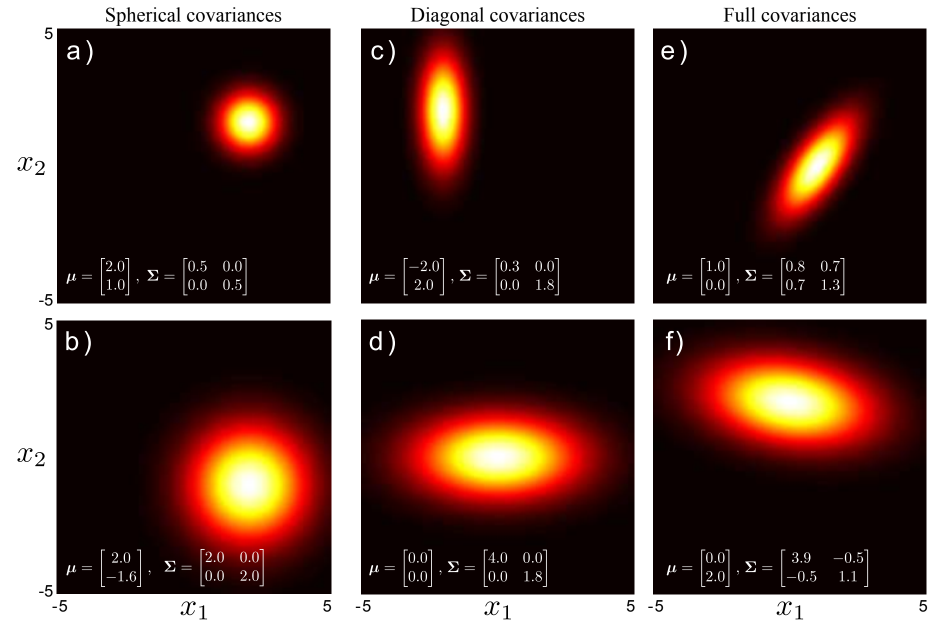

Figure 5.1 Covariance matrices take three forms. a-b) Spherical covariance matrices are multiples of the identity. The variables are independent and the iso-probability surfaces are hyperspheres. c-d) Diagonal covariance matrices permit different non-zero entries on the diagonal, but have zero entries elsewhere. The variables are independent, but scaled differently and the iso-probability surfaces are hyper-ellipsoids (ellipses in 2D) whose principal axes are aligned to the coordinate axes. e-f) Full covariance matrices are symmetric and positive definite. Variables are dependent and iso-probability surfaces are ellipsoids that are not aligned in any special way.

For the bivariate case (figure 5.1), spherical covariances produce circular iso-density contours. Diagonal covariances produce ellipsoidal iso-contours that are aligned with the coordinate axes. Full covariances also produce ellipsoidal iso-density contours, but these may now take an arbitrary orientation. More generally, in D dimensions, spherical covariances produce iso-contours that are D-spheres, diagonal covariances produce iso-contours that are D-dimensional ellipsoids aligned with the coordinate axes, and full covariances produce iso-contour that are n-dimensional ellipsoids in general position.

When the covariance is spherical or diagonal, the individual variables are independent. For example, for the bivariate diagonal case with zero mean, we have

$$\begin{aligned} Pr(x_1, x_2) &= \frac{1}{2\pi\sqrt{|\mathbf{\Sigma}|}} \exp\left[-0.5 \begin{pmatrix} x_1 & x_2 \end{pmatrix} \mathbf{\Sigma}^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right] \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-0.5 \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right] \end{aligned}$$

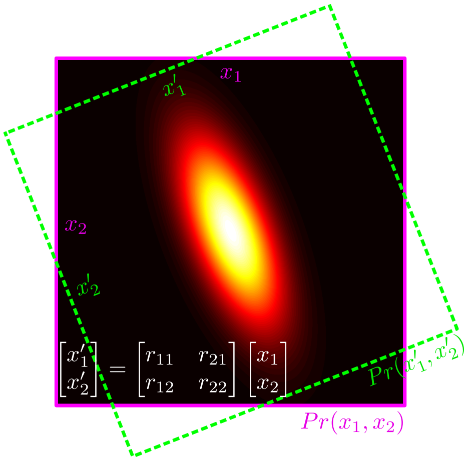
x'_1
x_1
x_2
x'_2
$\begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{21} \\ r_{12} & r_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
$Pr(x'_1, x'_2)$
$Pr(x_1, x_2)$

Figure 5.2 Decomposition of full covariance. For every bivariate normal distribution in variables x_1 and x_2 with full covariance matrix, there exists a coordinate system with variables x'_1 and x'_2 where the covariance is diagonal: the ellipsoidal iso-contours align with the coordinate axes x'_1 and x'_2 in this canonical coordinate frame. The two frames of reference are related by the rotation matrix $\mathbf{R}$ which maps (x'_1, x'_2) to (x_1, x_2). From this it follows (see text) that any covariance matrix $\mathbf{\Sigma}$ can be broken down into the product $\mathbf{R}^T\mathbf{\Sigma}'_{diag}\mathbf{R}$ of a rotation matrix $\mathbf{R}$ and a diagonal covariance matrix $\mathbf{\Sigma}'_{diag}$.

$$\begin{aligned} &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{x_1^2}{2\sigma_1^2}\right]\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left[-\frac{x_2^2}{2\sigma_2^2}\right] \\ &= Pr(x_1)Pr(x_2). \end{aligned} \tag{5.4}$$

5.2 Decomposition of covariance

We can use the foregoing geometrical intuitions to decompose the full covariance matrix $\mathbf{\Sigma}_{full}$. Given a normal distribution with mean zero and a full covariance matrix, we know that the iso-contours take an ellipsoidal form with the major and minor axes at arbitrary orientations.

Now consider viewing the distribution in a new coordinate frame where the axes *are* aligned with the axes of the normal (figure 5.2): in this new frame of reference, the covariance matrix $\mathbf{\Sigma}'_{diag}$ will be diagonal. We denote the data vector in the new coordinate system by $\mathbf{x}' = [x'_1, x'_2]^T$ where the frames of reference are related by $\mathbf{x}' = \mathbf{R}\mathbf{x}$. We can write the probability distribution over $\mathbf{x}'$ as

$$Pr(\mathbf{x}') = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}'_{diag}|^{1/2}}\exp\left[-0.5\mathbf{x}'^T\mathbf{\Sigma}'^{-1}_{diag}\mathbf{x}'\right]. \tag{5.5}$$

We now convert back to the original axes by substituting in $\mathbf{x}' = \mathbf{R}\mathbf{x}$ to get

$$\begin{aligned} Pr(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}'_{diag}|^{1/2}}\exp\left[-0.5(\mathbf{R}\mathbf{x})^T\mathbf{\Sigma}'^{-1}_{diag}\mathbf{R}\mathbf{x}\right] \\ &= \frac{1}{(2\pi)^{D/2}|\mathbf{R}^T\mathbf{\Sigma}'_{diag}\mathbf{R}|^{1/2}}\exp\left[-0.5\mathbf{x}^T(\mathbf{R}^T\mathbf{\Sigma}'_{diag}\mathbf{R})^{-1}\mathbf{x}\right] \end{aligned} \tag{5.6}$$

where we have used $|\mathbf{R}^T\mathbf{\Sigma}'\mathbf{R}| = |\mathbf{R}^T|.|\mathbf{\Sigma}'|.|\mathbf{R}| = 1.|\mathbf{\Sigma}'|.1 = |\mathbf{\Sigma}'|$. Equation 5.6 is a multivariate normal with covariance

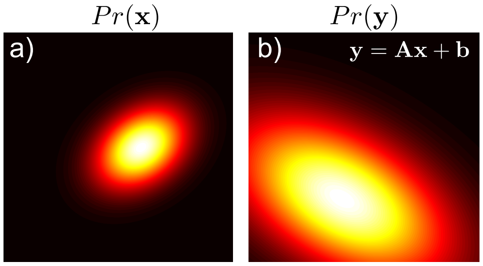

Figure 5.3 Transformation of normal variables. a) If $\mathbf{x}$ has a multivariate normal pdf and we apply a linear transformation to create new variable $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, then b) the distribution of $\mathbf{y}$ is also multivariate normal. The mean and covariance of $\mathbf{y}$ depend on the original mean and covariance of $\mathbf{x}$ and the parameters $\mathbf{A}$ and $\mathbf{b}$.

$$\boldsymbol{\Sigma}_{full} = \mathbf{R}^T\boldsymbol{\Sigma}'_{diag}\mathbf{R}. \tag{5.7}$$

We conclude that full covariance matrices are expressible as a product of this form involving a rotation matrix $\mathbf{R}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}'_{diag}$. Having understood this, it is possible to retrieve these elements from an arbitrary valid covariance matrix $\boldsymbol{\Sigma}_{full}$ by decomposing it in this way using the singular value decomposition.

The matrix $\mathbf{R}$ contains the principal directions of the ellipsoid in its columns. The values on the diagonal of $\boldsymbol{\Sigma}'_{diag}$ encode the variance (and hence the width of the distribution) along each of these axes. Hence we can use the results of the eigen-decomposition to answer questions about which directions in space are most and least certain.

5.3 Linear transformations of variables

Problem 5.1
Problem 5.2

The form of the multivariate normal is preserved under linear transformations $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ (figure 5.3). If the original distribution was

$$Pr(\mathbf{x}) = \text{Norm}_{\mathbf{x}}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}\right], \tag{5.8}$$

then the transformed variable $\mathbf{y}$ is distributed as:

$$Pr(\mathbf{y}) = \text{Norm}_{\mathbf{y}}\left[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\right]. \tag{5.9}$$

This relationship provides a simple method to draw samples from a normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We first draw a sample $\mathbf{x}$ from a standard normal distribution (with mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance $\boldsymbol{\Sigma} = \mathbf{I}$) and then apply the transform $\mathbf{y} = \boldsymbol{\Sigma}^{1/2}\mathbf{x} + \boldsymbol{\mu}$.

5.4 Marginal distributions

Problem 5.3

If we marginalize over any subset of random variables in a multivariate normal

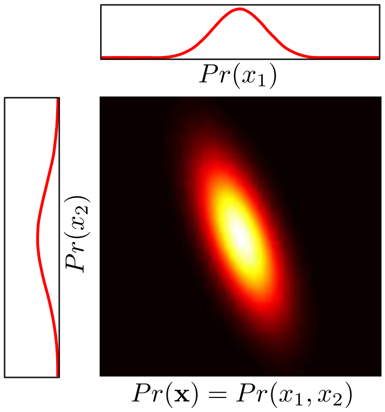

Figure 5.4 The marginal distribution of any subset of variables in a normal distribution is also normally distributed. In other words, if we sum over the distribution in any direction, the remaining quantity is also normally distributed. To find the mean and the covariance of the new distribution, we can simply extract the relevant entries from the original mean and covariance matrix.

distribution, the remaining distribution is also normally distributed (figure 5.4). If we partition the original random variable into two parts $\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T]^T$ so that

$$Pr(\mathbf{x}) = Pr\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\right) = \text{Norm}_{\mathbf{x}}\left[\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{21}^T\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right], \tag{5.10}$$

then

$$\begin{aligned} Pr(\mathbf{x}_1) &= \text{Norm}_{\mathbf{x}_1}[\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}] \\ Pr(\mathbf{x}_2) &= \text{Norm}_{\mathbf{x}_2}[\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}]. \end{aligned} \tag{5.11}$$

So, to find the mean and covariance of the marginal distribution of a subset of variables, we extract the relevant entries from the original mean and covariance.

5.5 Conditional distributions

If the variable $\mathbf{x}$ is distributed as a multivariate normal, then the conditional distribution of a subset of variables $\mathbf{x}_1$ given known values for the remaining variables $\mathbf{x}_2$ is also distributed as a multivariate normal (figure 5.5). If

Problem 5.4
Problem 5.5
Problem 5.6

$$Pr(\mathbf{x}) = Pr\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\right) = \text{Norm}_{\mathbf{x}}\left[\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{21}^T\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right], \tag{5.12}$$

then the conditional distributions are

$$\begin{aligned} Pr(\mathbf{x}_1|\mathbf{x}_2=\mathbf{x}_2^*) &= \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu_1}+\boldsymbol{\Sigma}_{21}^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2^*-\boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{21}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right] \quad (5.13)\\ Pr(\mathbf{x}_2|\mathbf{x}_1=\mathbf{x}_1^*) &= \text{Norm}_{\mathbf{x}_2}\left[\boldsymbol{\mu_2}+\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1^*-\boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{21}^T\right]. \end{aligned}$$

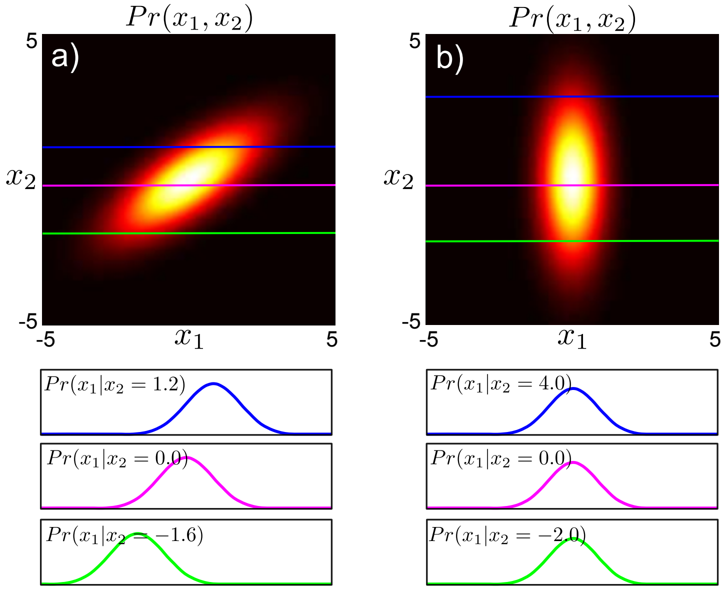

Figure 5.5 Conditional distributions of multivariate normal. a) If we take any multivariate normal distribution, fix a subset of the variables, and look at the distribution of the remaining variables, this distribution will also take the form of a normal. The mean of this new normal depends on the values that we fixed the subset to, but the covariance is always the same. b) If the original multivariate normal has spherical or diagonal covariance, both the mean and covariance of the resulting normal distributions are the same, regardless of the value we conditioned on: these forms of covariance matrix imply independence between the constituent variables.

5.6 Product of two normals

Problem 5.7
Problem 5.8

The product of two normal distributions is proportional to a third normal distribution (figure 5.6). If the two original distributions have means $\mathbf{a}$ and $\mathbf{b}$ and covariances $\mathbf{A}$ and $\mathbf{B}$, respectively, then we find that

$$\begin{aligned}\text{Norm}_{\mathbf{x}}[\mathbf{a},\mathbf{A}]\text{Norm}_{\mathbf{x}}[\mathbf{b},\mathbf{B}] &= \qquad (5.14)\\ &\kappa \cdot \text{Norm}_{\mathbf{x}}\left[\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}), \left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}\right],\end{aligned}$$

Problem 5.9

where the constant κ is itself a normal distribution,

$$\kappa = \text{Norm}_{\mathbf{a}}[\mathbf{b},\mathbf{A}+\mathbf{B}] = \text{Norm}_{\mathbf{b}}[\mathbf{a},\mathbf{A}+\mathbf{B}]. \qquad (5.15)$$

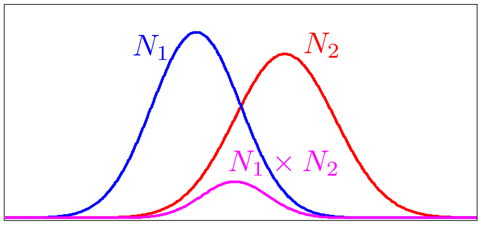

Figure 5.6 The product of any two normals N_1 and N_2 is proportional to a third normal distribution, with a mean between the two original means and a variance that is smaller than either of the original distributions.

5.6.1 Self-conjugacy

The preceding property can be used to demonstrate that the normal distribution is *self-conjugate* with respect to its mean $\boldsymbol{\mu}$. Consider taking a product of a normal distribution over data $\mathbf{x}$ and a second normal distribution over the mean vector $\boldsymbol{\mu}$ of the first distribution. It is easy to show from equation 5.14 that

$$\begin{aligned}\text{Norm}_{\mathbf{x}}[\boldsymbol{\mu},\boldsymbol{\Sigma}]\text{Norm}_{\boldsymbol{\mu}}[\boldsymbol{\mu}_p,\boldsymbol{\Sigma}_p] &= \text{Norm}_{\boldsymbol{\mu}}[\mathbf{x},\boldsymbol{\Sigma}]\text{Norm}_{\boldsymbol{\mu}}[\boldsymbol{\mu}_p,\boldsymbol{\Sigma}_p]\\ &= \kappa\cdot\text{Norm}_{\boldsymbol{\mu}}[\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}], \qquad (5.16)\end{aligned}$$

which is the definition of conjugacy (see section 3.9). The new parameters $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are determined from equation 5.14. This analysis assumes that the variance $\boldsymbol{\Sigma}$ is being treated as a fixed quantity. If we also treat this as uncertain, then we must use a normal inverse Wishart prior.

5.7 Change of variable

Consider a normal distribution in variable $\mathbf{x}$ whose mean is a linear function $\mathbf{Ay}+\mathbf{b}$ of a second variable $\mathbf{y}$. We can re-express this in terms of a normal distribution in $\mathbf{y}$ which is a linear function $\mathbf{A}'\mathbf{x}+\mathbf{b}'$ of $\mathbf{x}$ so that

$$\text{Norm}_{\mathbf{x}}[\mathbf{Ay}+\mathbf{b},\boldsymbol{\Sigma}] = \kappa\cdot\text{Norm}_{\mathbf{y}}[\mathbf{A}'\mathbf{x}+\mathbf{b}',\boldsymbol{\Sigma}'], \qquad (5.17)$$

where κ is a constant and the new parameters are given by

Problem 5.10

$$\begin{aligned}\boldsymbol{\Sigma}' &= (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\\ \mathbf{A}' &= (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\\ \mathbf{b}' &= -(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}. \qquad (5.18)\end{aligned}$$

This relationship is mathematically opaque, but it is easy to understand visually when x and y are scalars (figure 5.7). It is often used in the context of Bayes' rule where our goal is to move from $Pr(\mathbf{x}|\mathbf{y})$ to $Pr(\mathbf{y}|\mathbf{x})$.

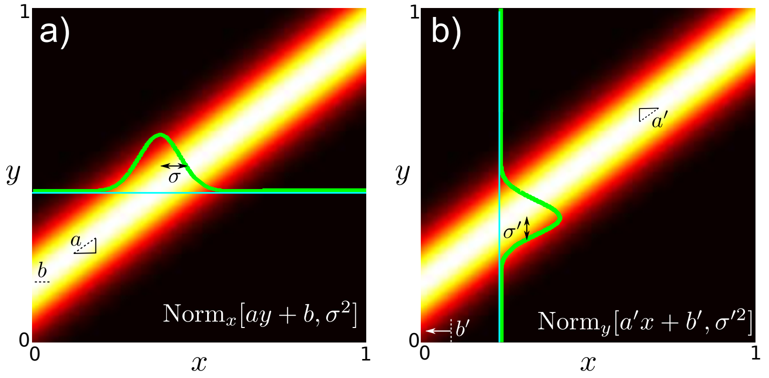

Figure 5.7 a) Consider a normal distribution in x whose variance σ^2 is constant, but whose mean is a linear function $ay + b$ of a second variable y. b) This is mathematically equivalent to a constant κ times a normal distribution in y whose variance σ'^2 is constant and whose mean is a linear function $a'x + b'$ of x.

Summary

In this chapter we have presented a number of properties of the multivariate normal distribution. The most important of these relate to the marginal and conditional distributions: when we marginalize or take the conditional distribution of a normal with respect to a subset of variables, the result is another normal. These properties are exploited in many vision algorithms.

Notes

The normal distribution has further interesting properties which are not discussed because they are not relevant for this book. For example, the convolution of a normal distribution with a second normal distribution produces a function that is proportional to a third normal, and the Fourier transform of a normal profile creates a normal profile in frequency space. For a different treatment of this topic the interested reader can consult chapter 2 of Bishop (2006).

Problems

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:

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

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 $\mathbf{y} = [\mathbf{I}, \mathbf{0}]\mathbf{x}$.

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.

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

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

Hint: use Schur's complement.

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).

Problem 5.7 Show that

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

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.

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

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

Problem 5.10 Prove the change of variable relation. Show that

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

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

Part II

Machine learning for machine vision

Part II: Machine learning for machine vision

In the second part of this book (chapters 6-9), we treat vision as a machine learning problem and disregard everything we know about the creation of the image. For example, we will not exploit our understanding of perspective projection or light transport. Instead, we treat vision as pattern recognition; we interpret new image data based on prior experience of images in which the contents were known. We divide this process into two parts: in *learning* we model the relationship between the image data and the scene content. In *inference*, we exploit this relationship to predict the contents of new images.

To abandon useful knowledge about image creation may seem odd, but the logic is twofold. First, these same learning and inference techniques will also underpin our algorithms when image formation is taken into account. Second, it is possible to achieve a great deal with a pure learning approach to vision. For many tasks, knowledge of the image formation process is genuinely unnecessary.

The structure of part II is as follows: in chapter 6 we present a taxonomy of models that relate the measured image data and the actual scene content. In particular, we distinguish between *generative* models and *discriminative* models. For generative models, we build a probability model of the data and parameterize it by the scene content. For discriminative models, we build a probability model of the scene content and parameterize it by the data. In the subsequent three chapters, we elaborate our discussion of these models.

In chapter 7 we consider generative models. In particular, we discuss how to use *hidden variables* to construct complex probability densities over visual data. As examples, we consider mixtures of Gaussians, t-distributions, and factor analyzers. Together, these three models allow us to build densities that are multi-modal, robust, and suitable for modeling high dimensional data.

In chapter 8 we consider *regression* models: we aim to estimate a continuous quantity from continuous data. For example, we might want to predict the joint angles from an image of the human body. We start with linear regression and move to more complex nonlinear methods such as Gaussian process regression and relevance vector regression. In chapter 9 we consider *classification* models: here we want to predict a discrete quantity from continuous data. For example, we might want to assign a label to a region of the image to indicate whether or not a face is present. We start with logistic regression and work toward more sophisticated methods such as Gaussian process classification, boosting, and classification trees.