Chapter 3

Common probability distributions

In chapter 2 we introduced abstract rules for manipulating probabilities. To use these rules we will need to define some probability distributions. The choice of distribution $Pr(x)$ that we use will depend on the *domain* of the data x that we are modeling (table 3.1).

Data Type	**Domain**	**Distribution**
univariate, discrete, binary	$x \in \{0,1\}$	Bernoulli
univariate, discrete, multi-valued	$x \in \{1,2,\ldots,K\}$	categorical
univariate, continuous, unbounded	$x \in \mathbb{R}$	univariate normal
univariate, continuous, bounded	$x \in [0,1]$	beta
multivariate, continuous, unbounded	$\mathbf{x} \in \mathbb{R}^K$	multivariate normal
multivariate, continuous, bounded, sums to one	$\mathbf{x} = [x_1, x_2, \ldots, x_K]^T$ $x_k \in [0,1], \sum_{k=1}^K x_k = 1$	Dirichlet
bivariate, continuous, x_1 unbounded, x_2 bounded below	$\mathbf{x} = [x_1, x_2]$ $x_1 \in \mathbb{R}$ $x_2 \in \mathbb{R}^+$	normal-scaled inverse gamma
multivariate vector $\mathbf{x}$ and matrix $\mathbf{X}$, $\mathbf{x}$ unbounded, $\mathbf{X}$ square, positive definite	$\mathbf{x} \in \mathbb{R}^K$ $\mathbf{X} \in \mathbb{R}^{K \times K}$ $\mathbf{z}^T\mathbf{X}\mathbf{z} > 0 \quad \forall\, \mathbf{z} \in \mathbb{R}^K$	normal inverse Wishart

Table 3.1: Common probability distributions: the choice of distribution depends on the type/domain of data to be modeled.

Probability distributions such as the categorical and normal distributions are obviously useful for modeling visual data. However, the need for some of the other

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

distributions is not so obvious; for example, the Dirichlet distribution models K positive numbers that sum to one. Visual data do not normally take this form.

The explanation is as follows: when we fit probability models to data, we need to know how uncertain we are about the fit. This uncertainty is represented as a probability distribution over the parameters of the fitted model. So for each distribution used for modeling, there is a second distribution over the associated parameters (table 3.2). For example, the Dirichlet is used to model the parameters of the categorical distribution. In this context, the parameters of the Dirichlet would be known as *hyperparameters*. More generally, the hyperparameters determine the shape of the distribution over the parameters of the original distribution.

Distribution	**Domain**	**Parameters modeled by**
Bernoulli	$x \in \{0, 1\}$	beta
categorical	$x \in \{1, 2, \ldots, K\}$	Dirichlet
univariate normal	$x \in \mathbb{R}$	normal inverse gamma
multivariate normal	$\mathbf{x} \in \mathbb{R}^k$	normal inverse Wishart

Table 3.2: Common distributions used for modeling (left) and their associated domains (center). For each of these distributions there is a second associated distribution over the parameters (right).

We will now work through the distributions in table 3.2 before looking more closely at the relationship between these pairs of distributions.

3.1 Bernoulli distribution

The *Bernoulli distribution* (figure 3.1) is a discrete distribution that models binary trials: it describes the situation where there are only two possible outcomes $x \in \{0, 1\}$ which are referred to as "failure" and "success." In machine vision, the Bernoulli distribution could be used to model the data. For example, it might describe the probability of a pixel taking an intensity value of greater or less than 128. Alternatively, it could be used to model the state of the world. For example, it might describe the probability that a face is present or absent in the image.

Problem 3.1

The Bernoulli has a single parameter $\lambda \in [0, 1]$ which defines the probability of observing a success $x = 1$. The distribution is hence

$$\begin{aligned} Pr(x=0) &= 1-\lambda \\ Pr(x=1) &= \lambda. \end{aligned} \tag{3.1}$$

We can alternatively express this as

$$Pr(x) = \lambda^x (1-\lambda)^{1-x}, \tag{3.2}$$

and we will sometimes use the equivalent notation

$$Pr(x) = \text{Bern}_x[\lambda]. \tag{3.3}$$

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

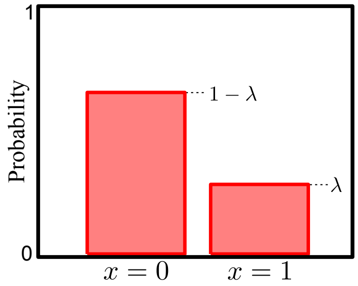


Figure 3.1 Bernoulli distribution. The Bernoulli distribution is a discrete distribution with two possible outcomes, $x \in \{0, 1\}$ which are referred to as failure and success, respectively. It is governed by a single parameter λ that determines the probability of success such that $Pr(x = 0) = 1 - \lambda$ and $Pr(x = 1) = \lambda$.

3.2 Beta distribution

The *beta distribution* (figure 3.2) is a continuous distribution defined on single variable λ where $\lambda \in [0, 1]$. As such it is suitable for representing uncertainty in the parameter λ of the Bernoulli distribution.

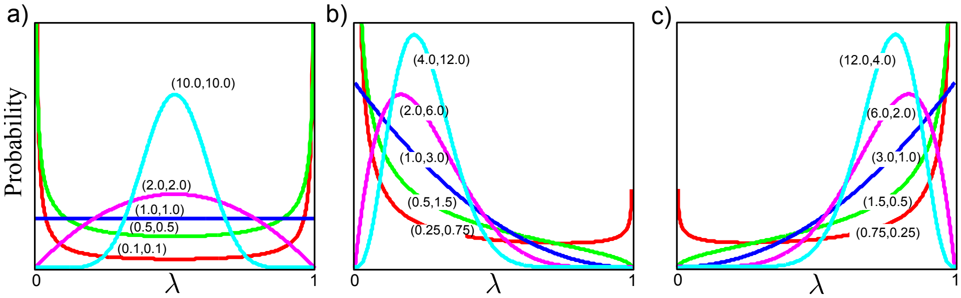


Figure 3.2 Beta distribution. The beta distribution is defined on $[0, 1]$ and has parameters (α, β) whose relative values determine the expected value so $E[\lambda] = \alpha/(\alpha + \beta)$ (numbers in parentheses show the α, β for each curve). As the absolute values of (α, β) increase, the concentration around $E[\lambda]$ increases. a) $E[\lambda] = 0.5$ for each curve, concentration varies. b) $E[\lambda] = 0.25$. c) $E[\lambda] = 0.75$.

The beta distribution has two parameters $\alpha, \beta \in [0, \infty]$ which both take positive values and affect the shape of the curve as indicated in figure 3.2. Mathematically, the beta distribution has the form

Problem 3.2
Problem 3.3
Problem 3.4

$$Pr(\lambda) = \frac{\Gamma[\alpha + \beta]}{\Gamma[\alpha]\Gamma[\beta]} \lambda^{\alpha-1}(1 - \lambda)^{\beta-1}, \tag{3.4}$$

where $\Gamma[\bullet]$ is the *gamma function*[1]. For short, we abbreviate this to

Problem 3.5

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

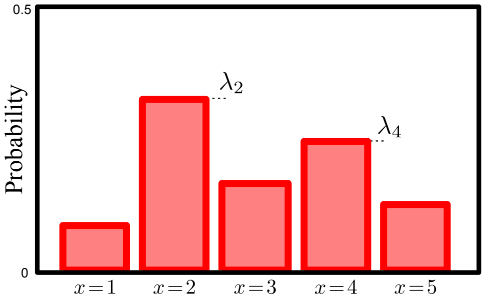


Figure 3.3 The categorical distribution is a discrete distribution with K possible outcomes, $x \in \{1, 2 \ldots, K\}$ and K parameters $\lambda_1, \lambda_2 \ldots, \lambda_K$ where $\lambda_k \geq 0$ and $\sum_k \lambda_k = 1$. Each parameter represents the probability of observing one of the outcomes, so that the probability of observing $x = k$ is given by λ_k. When the number of possible outcomes K is 2, the categorical reduces to the Bernoulli distribution.

$$Pr(\lambda) = \text{Beta}_\lambda[\alpha, \beta]. \tag{3.5}$$

3.3 Categorical distribution

The *categorical distribution* (figure 3.3) is a discrete distribution that determines the probability of observing one of K possible outcomes. Hence, the Bernoulli distribution is a special case of the categorical distribution when there are only two outcomes. In machine vision the intensity data at a pixel is usually quantized into discrete levels and so can be modeled with a categorical distribution. The state of the world may also take one of several discrete values. For example an image of a vehicle might be classified into {car,motorbike,van,truck} and our uncertainty over this state could be described by a categorical distribution.

The probabilities of observing the K outcomes are held in a $K \times 1$ parameter vector $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_K]$, where $\lambda_k \in [0, 1]$ and $\sum_{k=1}^{K} \lambda_k = 1$. The categorical distribution can be visualized as a normalized histogram with K bins and can be written as

$$Pr(x = k) = \lambda_k. \tag{3.6}$$

For short, we use the notation

$$Pr(x) = \text{Cat}_x\left[\boldsymbol{\lambda}\right]. \tag{3.7}$$

Alternatively, we can think of the data as taking values $\mathbf{x} \in \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$ where $\mathbf{e}_k$ is the k^{th} unit vector; all elements of $\mathbf{e}_k$ are zero except the k^{th}, which is one. Here we can write

$$Pr(\mathbf{x} = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{x_j} = \lambda_k, \tag{3.8}$$

where x_j is the j^{th} element of $\mathbf{x}$.

[1] The gamma function is defined as $\Gamma[z] = \int_0^\infty t^{z-1} e^{-t} dt$ and is closely related to factorials, so that for positive integers $\Gamma[z] = (z-1)!$ and $\Gamma[z+1] = z\Gamma[z]$.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

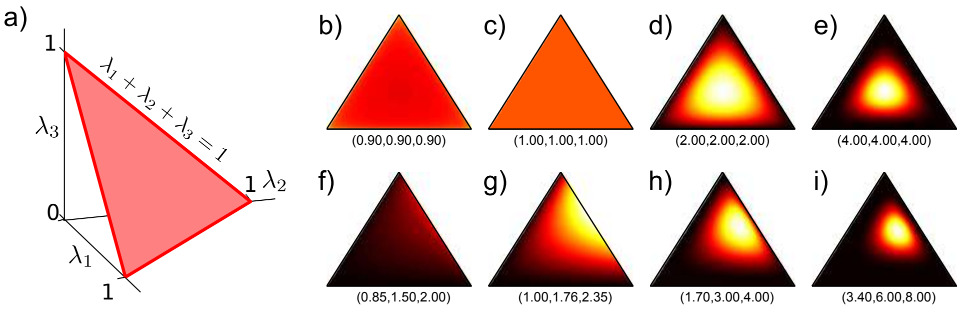


Figure 3.4 The Dirichlet distribution in K dimensions is defined on values $\lambda_1, \lambda_2, \ldots, \lambda_K$ such that $\sum_k \lambda_k = 1$ and $\lambda_k \in [0,1] \;\; \forall \;\; k \in \{1 \ldots K\}$. a) For K=3, this corresponds to a triangular section of the plane $\sum_k \lambda_k = 1$. In K dimensions, the Dirichlet is defined by K positive parameters $\alpha_{1 \ldots K}$. The ratio of the parameters determines the expected value for the distribution. The absolute values determine the concentration: the distribution is highly peaked around the expected value at high parameter values but pushed away from the expected value at low parameter values. b-e) Ratio of parameters is equal, absolute values increase. f-i) Ratio of parameters favors $\alpha_3 > \alpha_2 > \alpha_1$, absolute values increase.

3.4 Dirichlet distribution

The *Dirichlet distribution* (figure 3.4) is defined over K continuous values $\lambda_1 \ldots \lambda_K$ where $\lambda_k \in [0,1]$ and $\sum_{k=1}^{K} \lambda_k = 1$. Hence it is suitable for defining a distribution over the parameters of the categorical distribution.

In K dimensions the Dirichlet distribution has K parameters $\alpha_1 \ldots \alpha_K$ each of which can take any positive value. The relative values of the parameters determine the expected values $\mathrm{E}[\lambda_1] \ldots \mathrm{E}[\lambda_k]$. The absolute values determine the concentration around the expected value. We write

$$Pr(\lambda_{1 \ldots K}) \quad = \quad \frac{\Gamma[\sum_{k=1}^{K} \alpha_k]}{\prod_{k=1}^{K} \Gamma[\alpha_k]} \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1}, \tag{3.9}$$

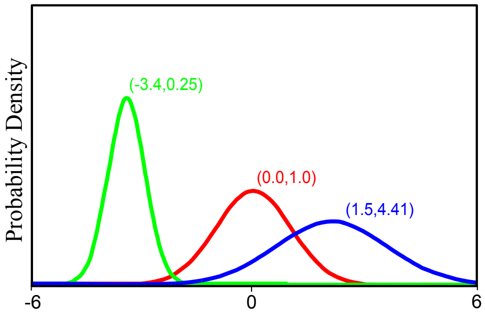
or for short

$$Pr(\lambda_{1 \ldots K}) = \mathrm{Dir}_{\lambda_{1 \ldots K}}[\alpha_{1 \ldots K}]. \tag{3.10}$$

Just as the Bernoulli distribution was a special case of the categorical distribution with two possible outcomes, so the beta distribution is a special case of the Dirichlet distribution where the dimensionality is two.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

Figure 3.5 The univariate normal distribution is defined on $x \in \mathbb{R}$ and has two parameters $\{\mu, \sigma^2\}$. The mean parameter μ determines the expected value and the variance σ^2 determines the concentration about the mean so that as σ^2 increases, the distribution becomes wider and flatter.



3.5 Univariate normal distribution

The *univariate normal* or *Gaussian distribution* (figure 3.5) is defined on continuous values $x \in [-\infty,\infty]$. In vision, it is common to ignore the fact that the intensity of a pixel is quantized and model it with the continuous normal distribution. The world state may also be described by the normal distribution. For example, the distance to an object could be represented in this way.

Problem 3.6
Problem 3.7

The normal distribution has two parameters, the mean μ and the variance σ^2. The parameter μ can take any value and determines the position of the peak. The parameter σ^2 takes only positive values and determines the width of the distribution. The normal distribution is defined as

$$Pr(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-0.5(x-\mu)^2/\sigma^2\right], \tag{3.11}$$

and we will abbreviate this by writing

$$Pr(x) = \text{Norm}_x[\mu, \sigma^2]. \tag{3.12}$$

3.6 Normal-scaled inverse gamma distribution

The *normal-scaled inverse gamma distribution* (figure 3.6) is defined over a pair of continuous values μ, σ^2, the first of which can take any value and the second of which is constrained to be positive. As such it can define a distribution over the mean and variance parameters of the normal distribution.

Problem 3.8

The normal-scaled inverse gamma has four parameters $\alpha, \beta, \gamma, \delta$ where α, β, and γ are positive real numbers but δ can take any value. It has pdf:

$$Pr(\mu, \sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma[\alpha]} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta-\mu)^2}{2\sigma^2}\right], \tag{3.13}$$

or for short

$$Pr(\mu, \sigma^2) = \text{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]. \tag{3.14}$$

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

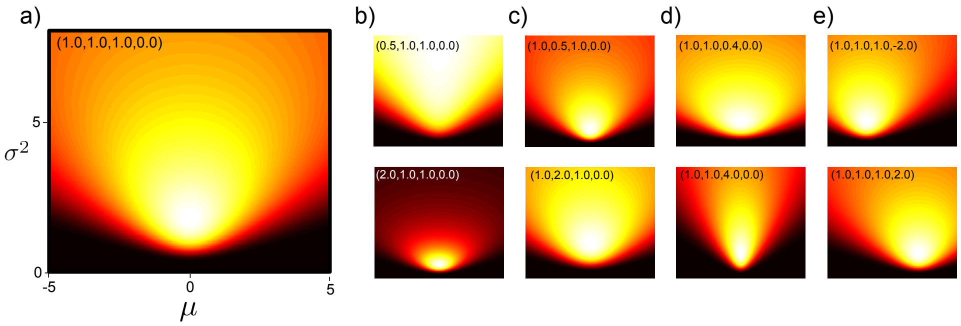


Figure 3.6 The normal-scaled inverse gamma distribution defines a probability distribution over bivariate continuous values μ, σ^2 where $\mu \in [-\infty, \infty]$ and $\sigma^2 \in [0, \infty]$. a) Distribution with parameters $[\alpha, \beta, \gamma, \delta] = [1, 1, 1, 0]$. b) Varying α. c) Varying β. d) Varying γ. e) Varying δ.

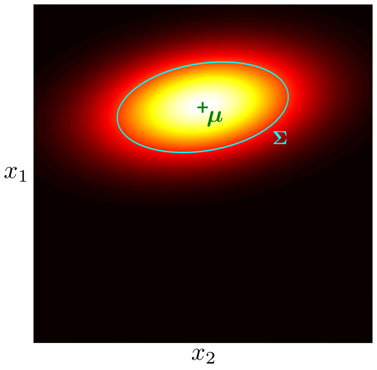


Figure 3.7 The multivariate normal distribution models D-dimensional variables $\mathbf{x} = [x_1 \ldots x_D]^T$ where each dimension x_d is continuous and real. It is defined by a $D \times 1$ vector $\boldsymbol{\mu}$ defining the mean of the distribution and a $D \times D$ covariance matrix $\boldsymbol{\Sigma}$ which determines the shape. The iso-contours of the distribution are ellipsoids where the center of the ellipsoid is determined by $\boldsymbol{\mu}$ and the shape by $\boldsymbol{\Sigma}$. This figure depicts a bivariate distribution, where the covariance is illustrated by drawing one of these ellipsoids.

3.7 Multivariate normal distribution

The *multivariate normal* or Gaussian distribution models D-dimensional variables $\mathbf{x}$ where each of the D elements $x_1 \ldots x_D$ is continuous and lies in the range $[-\infty, +\infty]$ (figure 3.7). As such the univariate normal distribution is a special case of the multivariate normal where the number of elements D is one. In machine vision the multivariate normal might model the joint distribution of the intensities of D pixels within a region of the image. The state of the world might also be described by this distribution. For example, the multivariate normal might describe the joint uncertainty in the 3D position (x, y, z) of an object in the scene.

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 mean of the distribution. The covariance $\boldsymbol{\Sigma}$ is a symmetric $D \times D$ positive definite matrix so that $\mathbf{z}^T\boldsymbol{\Sigma}\mathbf{z}$ is positive for any real vector $\mathbf{z}$. The probability density function has the following form

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

$$Pr(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-0.5(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \qquad (3.15)$$

or for short

$$Pr(\mathbf{x}) = \text{Norm}_{\mathbf{x}}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}\right]. \qquad (3.16)$$

The multivariate normal distribution will be used extensively throughout this book, and we devote the whole of chapter 5 to describing its properties.

3.8 Normal inverse Wishart distribution

The *normal inverse Wishart distribution* defines a distribution over a $D \times 1$ vector $\boldsymbol{\mu}$ and a $D \times D$ positive definite matrix $\boldsymbol{\Sigma}$. As such it is suitable for describing uncertainty in the parameters of a multivariate normal distribution. The normal inverse Wishart has four parameters $\alpha, \boldsymbol{\Psi}, \gamma, \boldsymbol{\delta}$, where α and γ are positive scalars, $\boldsymbol{\delta}$ is a $D \times 1$ vector and $\boldsymbol{\Psi}$ is a positive definite $D \times D$ matrix

$$Pr(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\gamma^{D/2}|\boldsymbol{\Psi}|^{\alpha/2}\exp\left[-0.5\left(\text{Tr}[\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}] + \gamma(\boldsymbol{\mu}-\boldsymbol{\delta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\delta})\right)\right]}{2^{\alpha D/2}(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{(\alpha+D+2)/2}\Gamma_D[\alpha/2]}, \qquad (3.17)$$

where $\Gamma_D[\bullet]$ is the multivariate gamma function and $\text{Tr}[\boldsymbol{\Psi}]$ returns the trace of the matrix $\boldsymbol{\Psi}$ (see appendix C.2.4). For short we will write:

$$Pr(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{NorIWis}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left[\alpha, \boldsymbol{\Psi}, \gamma, \boldsymbol{\delta}\right]. \qquad (3.18)$$

The mathematical form of the normal inverse Wishart distribution is rather opaque. However, it is just a function that produces a positive value for any valid mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, such that when we integrate over all possible values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the answer is one. It is hard to visualize the normal inverse Wishart, but easy to draw samples and examine them: each sample is the mean and covariance of a normal distribution (figure 3.8).

3.9 Conjugacy

We have argued that the beta distribution can represent probabilities over the parameters of the Bernoulli. Similarly the Dirichlet defines a distribution over the parameters of the categorical, and there are analogous relationships between the normal-scaled inverse gamma and univariate normal and the normal inverse Wishart and the multivariate normal.

These pairs were carefully chosen because they have a special relationship: in each case, the former distribution is *conjugate* to the latter: the beta is *conjugate* to the Bernoulli and the Dirichlet is conjugate to the categorical and so on. When

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

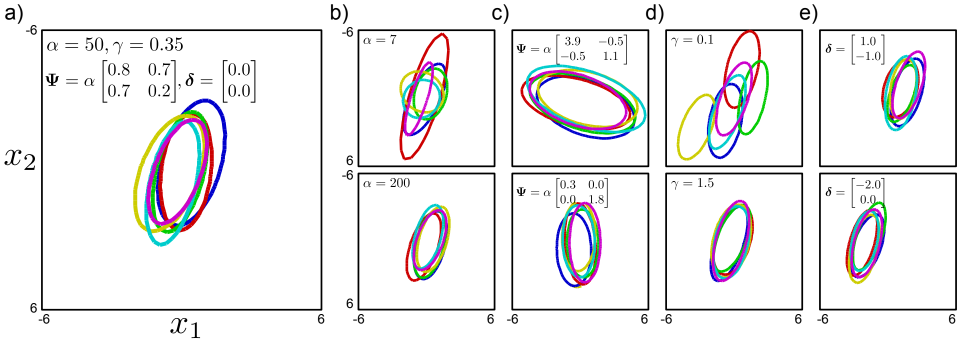


Figure 3.8 Sampling from 2D normal inverse Wishart distribution. a) Each sample consists of a mean vector and covariance matrix, here visualized with 2D ellipses illustrating the iso-contour of the associated Gaussian at a fixed Mahalanobis distance (i.e., a fixed proportion of the way through the density from the mean). b) Changing α modifies the dispersion of covariances observed. c) Changing $\mathbf{\Psi}$ modifies the average covariance. d) Changing γ modifies the dispersion of mean vectors observed. e) Changing $\boldsymbol{\delta}$ modifies the average value of the mean vectors.

we multiply a distribution with its conjugate, the result is proportional to a new distribution which has the same form as the conjugate. For example

Problem 3.9
Problem 3.10
Problem 3.11
Problem 3.12

$$\text{Bern}_x[\lambda] \cdot \text{Beta}_\lambda[\alpha, \beta] = \kappa(x, \alpha, \beta) \cdot \text{Beta}_\lambda\left[\tilde{\alpha}, \tilde{\beta}\right], \tag{3.19}$$

where κ is a scaling factor that is constant with respect to the variable of interest, λ. It is important to realize that this was not necessarily the case: if we had picked any distribution other than the beta, then this product would not have retained the same form. For this case, the relationship in equation 3.19 is easy to prove

$$\begin{aligned}
\text{Bern}_x[\lambda] \cdot \text{Beta}_\lambda[\alpha, \beta] &= \lambda^x (1-\lambda)^{1-x} \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]} \lambda^{\alpha-1}(1-\lambda)^{\beta-1} \\
&= \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]} \lambda^{x+\alpha-1}(1-\lambda)^{1-x+\beta-1} \\
&= \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]} \frac{\Gamma[x+\alpha]\Gamma[1-x+\beta]}{\Gamma[x+\alpha+1-x+\beta]} \text{Beta}_\lambda[x+\alpha, 1-x+\beta] \\
&= \kappa(x, \alpha, \beta) \cdot \text{Beta}_\lambda\left[\tilde{\alpha}, \tilde{\beta}\right].
\end{aligned} \tag{3.20}$$

where in the third line we have both multiplied and divided by the constant associated with $\text{Beta}_\lambda[\tilde{\alpha}, \tilde{\beta}]$.

The conjugate relationship is important because we take products of distributions during both learning (fitting distributions) and evaluating the model (assessing probability of new data under fitted distribution). The conjugate relationship means that these products can both be computed neatly in closed form.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

Summary

We use probability distributions to describe both the world state and the image data. We have presented four distributions (Bernoulli, categorical, univariate normal, multivariate normal) that are suited to this purpose. We also presented four other distributions (beta, Dirichlet, normal-scaled inverse gamma, and normal inverse Wishart) that can be used to describe the uncertainty in parameters of the first; they can hence describe the uncertainty in the fitted model. These four pairs of distributions have a special relationship: each distribution from the second set is conjugate to one from the first set. As we shall see, the conjugate relationship makes it easier to fit these distributions to observed data and evaluate new data under the fitted model.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

Notes

Throughout this book, I use rather esoteric terminology for discrete distributions. I distinguish between the *binomial distribution* (probability of getting M successes in N binary trials) and the *Bernoulli distribution* (the binary trial itself or probability of getting a success or failure in one trial) and talk exclusively about the latter distribution. I take a similar approach to discrete variables which can take K values. The *multinomial distribution* assigns a probability to observing the values {1,2,. . . , K} with frequency $\{M_1, M_2, \ldots, M_K\}$ given N trials. The *categorical distribution* is a special case of this with $N = 1$. Most other authors do not make this distinction and would term this 'multinomial' as well.

A more complete list of common probability distributions and details of their properties are given in Appendix B of Bishop (2006). Further information about conjugacy can be found in Chapter 2 of Bishop (2006) or any textbook on Bayesian methods, such as that of Gelman *et al.* (2004). Much more information about the normal distribution is provided in chapter 5 of this book.

Problems

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

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

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

Problem 3.5 Use integration by parts to prove that if

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

then

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

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

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.

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

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

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

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.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

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

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

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 $\text{Tr}\left[\mathbf{z}\mathbf{z}^T\mathbf{A}^{-1}\right]=\mathbf{z}^T\mathbf{A}^{-1}\mathbf{z}$.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.