Part I

Probability

Part I: Probability

We devote the first part of this book (chapters 1-4) to a brief review of probability and probability distributions. Almost all models for computer vision can be interpreted in a probabilistic context, and in this book we will present all the material in this light. The probabilistic interpretation may initially seem confusing, but it has a great advantage: it provides a common notation which will be used throughout the book and elucidates relationships between different models that would otherwise remain opaque.

So why is probability a suitable language to describe computer vision problems? In a camera the three-dimensional world is projected onto the optical surface to form a two-dimensional set of measurements. Our goal is to take these measurements and use them to establish the properties of the world that created them. However, there are two problems. First, the measurement process is noisy: what we observe is not the amount of light that fell on the sensor, but a noisy estimate of this quantity. We must describe the noise in this data and for this we use probability. Second, the relationship between world and measurements is generally many to one: there may be many real world configurations that produce exactly the same measurements. The relative likelihood of these possible worlds can also be described using probability.

The structure of part I is as follows: in chapter 1 we introduce the basic rules for manipulating probability distributions including the ideas of conditional and marginal probability and Bayes' rule. We also introduce more advanced ideas such as independence, conditional independence and expectation.

In chapter 2 we discuss the properties of eight specific probability distributions. We divide these into four pairs. The first set will be used to describe the observed data or the world. The second set of distributions model the parameters of the first set. In combination, they allow us to fit a probability model and provide information about how certain we are about the fit.

In chapter 3 we discuss methods for fitting probability distributions to observed data. We also discuss how to assess the probability of new data points under the fitted model and in particular, how to take account of uncertainty in the original fit when we calculate this predictive density. Finally, in chapter 4 we investigate the properties of the multivariate normal distribution in detail. This distribution is ubiquitous in vision applications and has a number of useful properties that are frequently exploited in machine vision.

Readers who are very familiar with probability models and the Bayesian philosophy may wish to skip this part and move directly to part II.

Chapter 1

Introduction to Probability

In this chapter, we provide a compact review of probability theory. There are very few ideas and each is relatively simple when considered separately. However, they combine to form a powerful language for describing uncertainty.

1.1 Random Variables

A random variable X denotes a quantity that is uncertain. The variable may denote the result of an experiment (e.g. flipping a coin) or a real-world measurement of a fluctuating property (e.g. measuring the temperature). If we observe several instances of X then it might take a different value on each occasion. However, some values may occur more often than others. This information is captured by the probability distribution $Pr(X)$ of the random variable.

A random variable may be *discrete* or *continuous*. A discrete variable takes values from a predefined set. This set may be ordered (the outcomes 1-6 of rolling a die) or unordered (the outcomes "sunny", "raining", "snowing" of observing the weather). It may be finite (there are 52 possible outcomes of drawing a card randomly from a standard pack) or infinite (the number of people on the next train is theoretically unbounded). The probability distribution of a discrete variable can be visualized as a histogram or a Hinton diagram (figure 1.1). Each outcome has a positive probability associated with it and the sum of the probabilities for all outcomes is always one.

Continuous random variables take values that are real numbers. These may be finite (the time taken to finish a 2 hour exam is constrained to be be greater than 0 hours and less than 2 hours) or infinite (the amount of time until the next bus arrives is unbounded above). Infinite continuous variables may be defined on the whole real range or may be bounded above or below (the velocity of a vehicle may take any value, but the speed is bounded below by 0). The probability distribution of a continuous variable can be visualized by plotting the probability density function (pdf). The probability density for an outcome represents the relative propensity of the random variable to take that value (see figure 1.2). It may take any positive value. However, the integral of the pdf always sums to one.

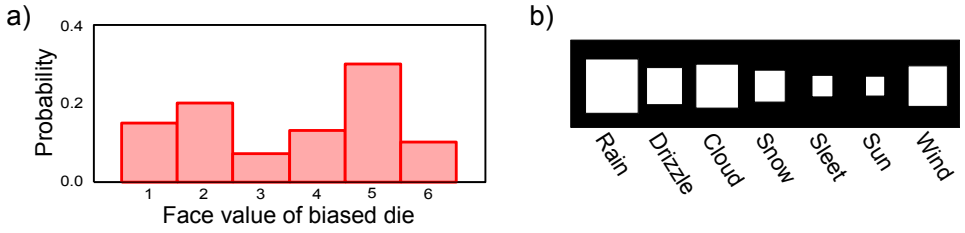

Figure 1.1 Two different representations for discrete probabilities a) A bar graph representing the probability that a biased 6-sided die lands on each face. The height of the bar represents the probability so the sum of all heights is one. b) A Hinton diagram illustrating the probability of observing different weather types in England. The area of the square represents the probability, so the sum of all areas is one.

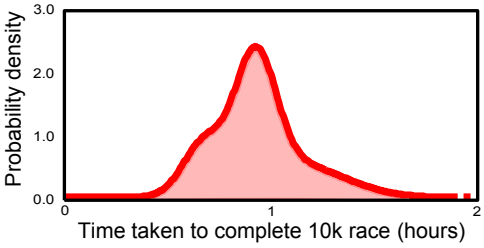

Figure 1.2 Continuous probability distribution (probability density function or PDF) for time taken for athletes to complete 10K race. Note that the probability density can exceed one, but the area under the curve must always have unit area.

1.2 Joint probability

Consider two random variables, X and Y. If we observe multiple paired instances of X and Y, then some combinations of the two outcomes occur more frequently than others. This information is encompassed in the *joint* probability distribution of X and Y which is written as $Pr(X, Y)$. The comma in $Pr(X, Y)$ can be read as the English word "and" so $Pr(X, Y)$ is the probability of X *and* Y. A joint probability distribution may relate variables that are all discrete, all continuous or it may relate discrete variables to continuous ones (see figure 1.3). Regardless, the total probability of all outcomes (summing over discrete variables and integrating over continuous ones), is always one.

In general we will be interested in the joint probability distribution of more than two variables. We will write $Pr(X, Y, Z)$ to represent the joint probability distribution of scalar variables X, Y, Z. We may also write $Pr(\mathbf{X})$ to represent the joint probability of all of the elements $X_1, X_2 \ldots X_K$ of the multidimensional variable $\mathbf{X}$. Finally, we will write $Pr(\mathbf{X}, \mathbf{Y})$ to represent the joint distribution of all of the elements from multidimensional variables $\mathbf{X}$ and $\mathbf{Y}$.

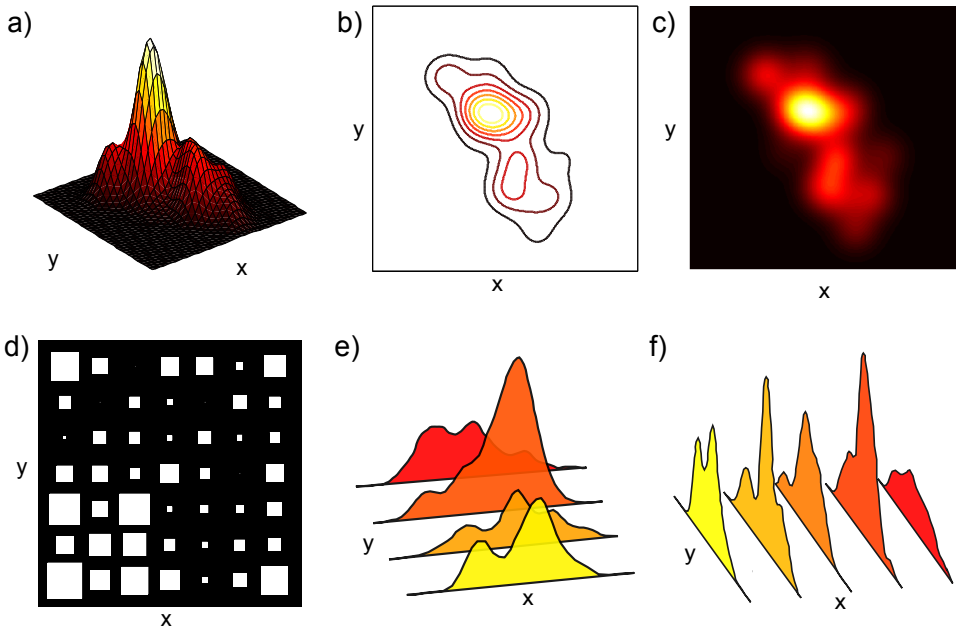

Figure 1.3 Joint probability distributions between variables X and Y. a-c) Joint pdf of two continuous variables represented as surface, contour plot and image respectively. d) Joint distribution of two discrete variables represented as 2d Hinton diagram e) Joint distribution of a continuous variable X and discrete variable Y. f) Joint distribution of a discrete variable X and continuous variable Y.

1.3 Marginalization

We can recover the probability distribution of any variable from a joint distribution by summing (discrete) or integrating (continuous) over all the other variables (figure 1.4). For example, if X and Y are both continuous and we know $Pr(X,Y)$, then we can recover the distributions $Pr(X)$ and $Pr(Y)$ using the relation:

$$Pr(X) = \int Pr(X,Y)dY \tag{1.1}$$

$$Pr(Y) = \int Pr(X,Y)dX \tag{1.2}$$

The recovered distributions $Pr(X)$ and $Pr(Y)$ are referred to as *marginal* distributions and the process of integrating/summing over the other variables is called *marginalization*. Calculating the marginal distribution $Pr(X)$ from the joint distribution $Pr(X,Y)$ by marginalizing over the variable Y has a simple interpretation: we are finding the probability distribution of X regardless of (or in the absence of information about) the value of Y.

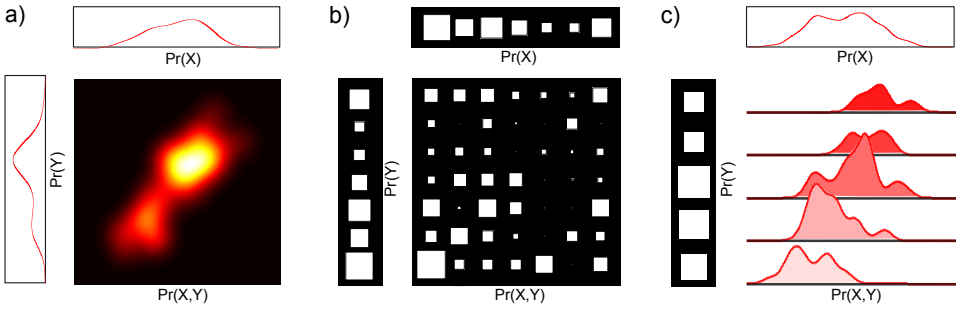

Figure 1.4 Joint and marginal probability distributions. The marginal probability $Pr(X)$ of X is found by summing over all values of Y in the joint distribution $Pr(X, Y)$ and vice-versa. Note that the plots for the marginal distributions have different scales from those for the joint distribution (on the same scale, they marginals would look larger as they sum all of the mass from one direction). a) Both X and Y are continuous. b) Both X and Y are discrete. c) The random variable X is continuous and Y is discrete.

In general, we can recover the joint probability of any subset of variables, by marginalizing over all of the others. For example, given four discrete variables, W, X, Y, Z we can recover $Pr(X, Y)$ using:

$$Pr(X, Y) = \sum_{W} \sum_{Z} Pr(W, X, Y, Z) \tag{1.3}$$

1.4 Conditional probability

The conditional probability of X given that Y takes value y^* tells us the relative propensity of the random variable X to take different outcomes given that the random variable Y is fixed to value y^*. This conditional probability is written as $Pr(X|Y = y^*)$. The vertical line "|" can be read as "given".

The conditional probability $Pr(X|Y = y^*)$ of X given that Y takes the value y^* can be recovered from the joint distribution $Pr(X, Y)$. In particular, we examine the appropriate slice $Pr(X, Y = y^*)$ of the joint distribution Pr(X,Y) (figure 1.5). The values in the slice tell us about the relative propensity of X to take various outcomes having observed $Y = y^*$ but do not themselves form a valid probability distribution: they cannot sum to one as they constitute only a small part of the joint distribution which does sum to one. To calculate the conditional probability distribution, we normalize by the total probability of this slice:

$$Pr(X|Y = y^*) = \frac{Pr(X, Y = y^*)}{\int Pr(X, Y = y^* dX} = \frac{Pr(X, Y = y^*)}{Pr(Y = y^*)} \tag{1.4}$$

where we have used the marginal probability relation (Equation 1.2) to simplify the

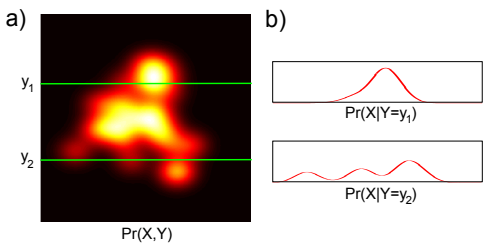

Figure 1.5 Conditional Probability. a) Joint pdf of X and Y. b) Two conditional probability distributions $Pr(X|Y = y_1)$ and $Pr(X|Y = y_2)$. These are formed by extracting the appropriate slice from the joint pdf and normalizing so that the area is one. A similar operation can be performed for discrete distributions.

denominator. It is common to write the conditional probability relation without explicitly defining the value $Y = y^*$ to give the more compact notation,

$$Pr(X|Y) = \frac{Pr(X,Y)}{Pr(Y)}. \tag{1.5}$$

This relationship can be re-arranged to give,

$$Pr(X,Y) = Pr(X|Y)Pr(Y), \tag{1.6}$$

and by symmetry,

$$Pr(X,Y) = Pr(Y|X)Pr(X). \tag{1.7}$$

When we have more than two variables, we may repeatedly take conditional probabilities to divide up the joint probability distribution into a product of terms:

$$\begin{aligned} Pr(W,X,Y,Z) &= Pr(W,X,Y|Z)Pr(Z) \\ &= Pr(W,X|Y,Z)Pr(Y|Z)Pr(Z) \\ &= Pr(W|X,Y,Z)Pr(X|Y,Z)Pr(Y|Z)Pr(Y) \end{aligned} \tag{1.8}$$

1.5 Bayes' rule

In equations 1.6 and 1.7 we expressed the joint probability in two ways. We can combine these formulations to find a relationship between $Pr(X|Y)$ and $Pr(Y|X)$:

$$Pr(Y|X)Pr(X) = Pr(X|Y)Pr(Y) \tag{1.9}$$

or rearranging

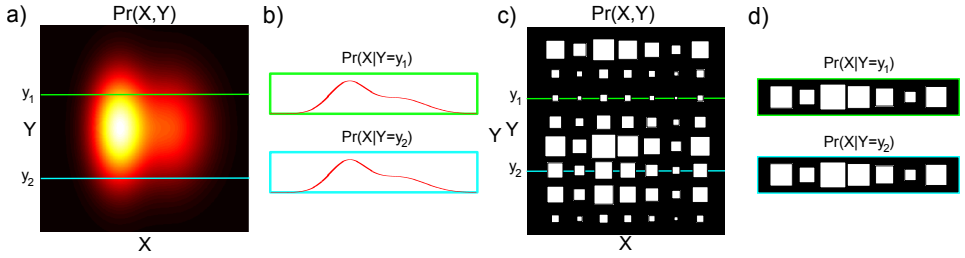

Figure 1.6 Independence. a) Joint pdf of continuous independent variables X and Y. b) The independence of X and Y means that every conditional distribution is the same: the value of Y tells us nothing about X and vice-versa. Compare this to figure 1.5 which illustrated variables that were dependent. c) Joint pdf of discrete independent variables X and Y. d) The conditional distributions of X given Y are all the same.

$$Pr(Y|X) = \frac{Pr(X|Y)Pr(Y)}{Pr(X)} \tag{1.10}$$

$$= \frac{Pr(X|Y)Pr(Y)}{\int Pr(X,Y)dY} \tag{1.11}$$

$$= \frac{Pr(X|Y)Pr(Y)}{\int Pr(X|Y)Pr(Y)dY} \tag{1.12}$$

where we have expanded the denominator in Equations 1.11 and 1.12 using the definitions of marginal and conditional probability respectively. Equations 1.10-1.12 are all commonly referred to as *Bayes' rule*.

Each term in Bayes's rule has a name. The term $Pr(Y|X)$ on the left hand side is the *posterior*. It represents what we know about Y given X. Conversely, the term $Pr(Y)$ is the *prior* as it represents what is known about Y before we know X. The term $Pr(X|Y)$ is the *likelihood* and the denominator $Pr(X)$ is the *evidence*.

1.6 Independence

If knowing the value of variable X tells us nothing about variable Y (and vice-versa) then we say X and Y are independent (figure 1.6). In this case, we can write:

$$Pr(X|Y) = Pr(X) \tag{1.13}$$

$$Pr(Y|X) = Pr(Y) \tag{1.14}$$

Substituting into equation 1.6 we see that for independent variables the joint probability Pr(X,Y) is the product of the marginal probabilities Pr(X) and Pr(Y).

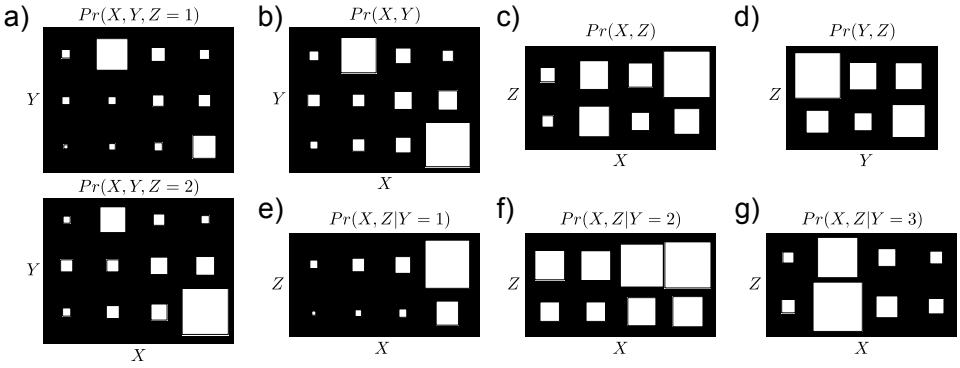

Figure 1.7 Conditional Independence. a) Joint pdf of three discrete variables X, Y, Z. All 24 probability values sum to one. b) Marginalizing, we see that variables X and Y are not independent - the conditional distribution of X is different for different values of Y and vice-versa. c) Variables X and Z are also dependent. d) Variables Y and Z are also dependent. e-g) However, X and Z are conditionally independent *given* Y. For fixed Y, X tells us nothing more about Z and vice-versa.

$$\begin{aligned} Pr(X,Y) &= Pr(X|Y)Pr(Y) \\ &= Pr(X)Pr(Y) \end{aligned} \tag{1.15}$$

This can be observed in figure 1.6: the joint distribution is the outer product of the two marginal distributions.

1.7 Conditional Independence

With more than two random variables, independence relations become more complex. The variable X is said to be *conditionally independent to variable Z given variable Y* when X and Z are independent for fixed Y (figure 1.7).

Confusingly, the conditional independence of X and Z given Y does not imply that X and Z are themselves independent. It merely implies that if we know variable Y then X provides no further information about Z and vice-versa. One way that this can occur is in a chain of events: if event X causes event Y and Y causes Z then the dependence of Z on X might be entirely mediated by Y.

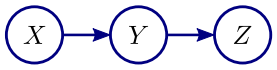

Figure 1.8 Directed graphical model relating variables X, Y, Z from figure 1.7. This model implies that the joint probability can be broken down as $Pr(X, Y, Z) = Pr(X)Pr(Y|X)Pr(Z|Y)$.

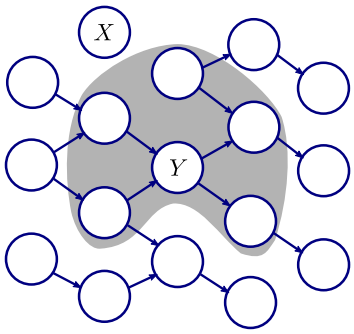

Figure 1.9 Interpreting directed graphical models. Variable X is independent from all the others as it is not connected to them. Variable Y is conditionally independent of all of the others given its *Markov blanket*. This comprises, the parents, children and other parents of the children. The Markov blanket for variable Y is indicated by the gray region.

1.8 Graphical Models

A graphical model is a graph-based representation of the dependencies between multiple variables. In a *directed graphical model* or *Bayesian network*, the dependencies are expressed as a factorization of the joint probability distribution into a product of conditional distributions so that

$$Pr(X_1 \ldots X_n) = \prod_{i=1}^{n} Pr(X_i|X_{pa[i]}), \tag{1.16}$$

where the notation $pa[i]$ denotes the set of variables that are parents of variable X_i. For example, the graphical model in figure 1.8 tells us that

$$Pr(X, Y, Z) = Pr(X)Pr(Y|X)Pr(Z|Y). \tag{1.17}$$

The conditional independence relations can be inferred from the visual representation of the directed graphical model by following two rules (figure 1.9). First, if there is no route connecting two variables at all then they are independent. Second, any variable is conditionally independent of all the other variables given its parents, children and the other parents of its children (its *Markov blanket*). Consequently, in figure 1.8, which corresponds to the conditional independence example above, variable X is conditionally independent of variable Z given variable Y.

Directed models will be used frequently throughout this book, but other types of graphical model also exist. For example, in chapter 11 we introduce undirected graphical models in which there is no notion of child and parent and different conditional independence relations are specified.

1.9 Expectation

Given a function $f()$ that returns a value for each possible value of X and a probability $Pr(X = x)$ that each value of X occurs, we sometimes wish to calculate the *expected* output of the function. If we drew a very large number of samples from the probability distribution, calculated the function for each sample and took the average of these values, the result would be the *expectation*. More precisely, the expected value of a function $f()$ of a random variable X is defined as

$$E[f(X)] = \sum_x f(x)Pr(X = x) \tag{1.18}$$

$$E[f(X)] = \int f(x)Pr(X = x)dx \tag{1.19}$$

for the discrete and continuous cases respectively. This idea generalizes to functions $f()$ of more than one random variable so that for example

$$E[f(X,Y)] = \iint f(x,y)Pr(X = x, Y = y)dxdy. \tag{1.20}$$

For some choices of the function $f()$ the expectation is given a special name as (table 1.1). Such quantities are are commonly used to summarize the properties of complex probability distributions. There are four rules for manipulating expectations, which can be easily proved from the original definition (equation 1.19).

$$E[k] = k \tag{1.21}$$

$$E[kf(X)] = kE[f(X)] \tag{1.22}$$

$$E[f(X) + g(X)] = E[f(X)] + E[g(X)] \tag{1.23}$$

$$E[f(X)g(X)] = E[f(X)]E[g(X)] \quad \text{if } X, Y \text{ independent} \tag{1.24}$$

Function f()	Expectation
x	mean, μ_x
x^k	k'th moment about zero
$(x-\mu_x)^k$	k'th moment about the mean
$(x-\mu_x)^2$	variance
$(x-\mu_x)^3$	skew
$(x-\mu_x)(y-\mu_y)$	covariance of X and Y

Table 1.1: Special cases of expectation. For some special functions $f(x)$, the expectation $E[f(x)]$ is given a special name. Here we use the notation μ_x to represent the mean with respect to random variable x and μ_y the mean with respect to random variable y.

Summary

The rules of probability are remarkably compact and simple. Amazingly, the ideas of marginalization, joint and conditional probability, independence and Bayes' rule will underpin all of the machine vision algorithms in this book.

Chapter 2

Common probability distributions

In chapter 1 we introduced abstract rules for manipulating probabilities. To use these rules we need mathematical expressions for probability distributions. The particular choices of expression $Pr(x)$ that we use will depend on the type of data x that we are modelling (table 2.1).

Data Type	**Domain**	**Distribution**
univariate, discrete, binary	$z \in \{0, 1\}$	Bernoulli
univariate, discrete, multi-valued	$z \in \{1, 2, \ldots, K\}$	categorical
univariate, continuous, unbounded	$z \in \mathbb{R}$	univariate normal
univariate, continuous, bounded	$z \in [0, 1]$	beta
multivariate, continuous, unbounded	$\mathbf{z} \in \mathbb{R}^K$	multivariate normal
multivariate, continuous, bounded, sums to one	$\mathbf{z} = [z_1 \ldots z_K]^T$ $z_k \in [0, 1], \sum_{k=1}^{K} z_k = 1$	Dirichlet
bivariate, continuous, z_1 unbounded, z_2 bounded below	$\mathbf{z} = [z_1, z_2]$ $z_1 \in \mathbb{R}$ $z_2 \in \mathbb{R}^+$	normal inverse gamma
multivariate vector $\mathbf{z}$ and matrix $\mathbf{Z}$ z unbounded, $\mathbf{Z}$ square, positive definite	$\mathbf{z} \in \mathbb{R}^k$ $\mathbf{Z} \in \mathbb{R}^{k \times k}$ $\mathbf{x}^T\mathbf{Z}\mathbf{x} > 0 \quad \forall \mathbf{x} \in \mathbb{R}^k$	normal inverse Wishart

Table 2.1: 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 distributions over more elaborate quantities is not so obvious: for example, the Dirichlet distribution

models K positive numbers that sum to one. It is hard to imagine visual data having this form.

The reason for these more elaborate distributions is as follows: when we fit probability models to data, we will need to know how uncertain we are about the fit. This uncertainty is represented as a probability distribution over the possible parameters of the fitted model. So for each distribution used for modelling, there is a second distribution over the associated parameters (table 2.2). In fact, the Dirichlet is used to model the parameters of the categorical distribution.

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

Table 2.2: (left) Common distributions used for modelling and (center) their associated domains. (right) For each of these distributions there is an associated distribution over the parameters.

We will now work through the distributions in table 2.2 in row order before looking more closely at the relationship between these pairs of distributions.

2.1 Bernoulli distribution

The Bernoulli distribution (figure 2.1) is a discrete distribution that models binary trials: it describes the situation where there are only two possible outcomes $y \in \{0, 1\}$ which are referred to as "failure" and "success". In machine vision, the Bernoulli distribution could be used to model the data: 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.

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

$$\begin{aligned} Pr(y = 0) &= 1 - \lambda \\ Pr(y = 1) &= \lambda. \end{aligned} \tag{2.1}$$

We can alternatively express this as

$$Pr(y) = \lambda^y (1 - \lambda)^{1-y}, \tag{2.2}$$

and we will sometimes use the equivalent notation

$$Pr(y) = \text{Bern}_y[\lambda]. \tag{2.3}$$

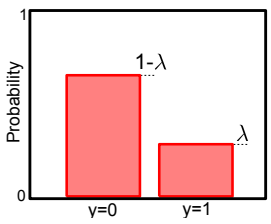

Figure 2.1 The Bernoulli distribution is a discrete distribution with two possible outcomes, $y \in \{0, 1\}$ referred to as failure and success repsectively. It is governed by a single parameter λ that determines the probability of success such that $Pr(y = 0) = 1 - \lambda$ and $Pr(y = 1) = \lambda$.

2.2 Beta distribution

The beta distribution (figure 2.2) is a continuous distribution defined on single parameter λ where $\lambda \in [0, 1]$. As such it is suitable for representing the uncertainty over the parameter λ of the Bernoulli distribution.

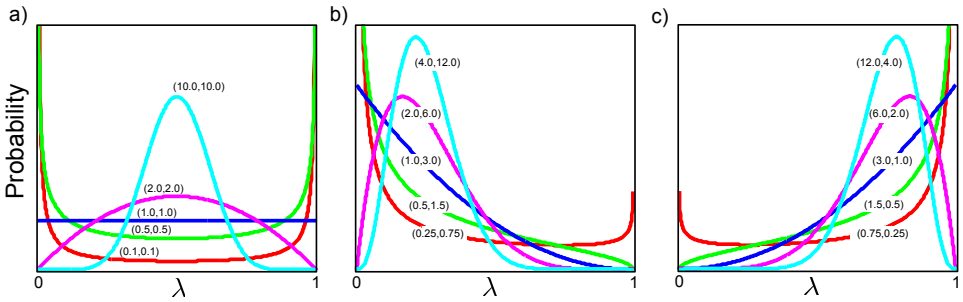

Figure 2.2 The Beta distribution is defined on $[0, 1]$ and has parameters (α, β) whose relative values determine the expected value so $E[\lambda] = \alpha/(\alpha + \beta)$. As the absolute values of (α, β) increase the concentration around $E[\lambda]$ increases. a) $E[\lambda] = 0.5$ for each curve, 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 effect the shape of the curve as indicated in figure 2.2. Mathematically, the beta distribution has the form,

$$Pr(\lambda) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \lambda^{\alpha-1}(1 - \lambda)^{\beta-1}. \tag{2.4}$$

where $\Gamma()$ is the Gamma function[1]. For short, we abbreviate this to

$$Pr(\lambda) = \text{Beta}_\lambda[\alpha, \beta]. \tag{2.5}$$

[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 $\Gamma(z) = (z - 1)!$

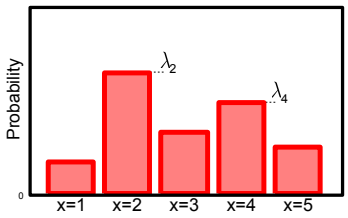

Figure 2.3 The categorical distribution is a discrete distribution with K possible outcomes, $x \in \{1, 2 \ldots K\}$ and K parameters $\lambda_1, \lambda_2 \ldots \lambda_K$ where $\sum_k \lambda_k = 1$. The likelihood of observing $x = k$ is given by λ_k. When K=2, the categorical reduces to the Bernoulli distribution.

2.3 Categorical distribution

The categorical distribution 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 modelled 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 K parameters $\lambda_1 \ldots \lambda_K$, 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

$$Pr(x = k) = \lambda_k. \tag{2.6}$$

Alternatively, we can think of the data as a vector $\mathbf{x} = [0, 0, \ldots 0, 1, 0, \ldots 0]$ where all elements are zero except the k'th which is one. Here we can write

$$Pr(\mathbf{x}) = \prod_{j=1}^{K} \lambda_j^{x_j} = \lambda_k, \tag{2.7}$$

where x_k is the k'th element of $\mathbf{x}$. For short, we use the notation

$$Pr(\mathbf{x}) = \mathrm{Cat}_x\left[\lambda_1 \ldots \lambda_K\right]. \tag{2.8}$$

2.4 Dirichlet distribution

The Dirichlet distribution (figure 2.4) 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 $E[\lambda_1] \ldots E[\lambda_k]$. The absolute values determine the concentration around the expected value. We write

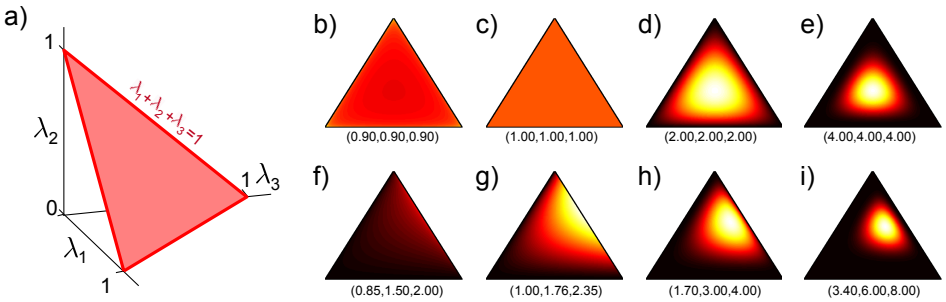

Figure 2.4 The Dirichlet distribution in K dimensions is defined on values $\lambda_1, \lambda_2 \ldots \lambda_K$ such that $\sum_k \lambda_k = 1$ and $\lambda_k \in [0,1] \;\forall\; k \in \{1 \ldots K\}$. a) For K=3, this corresponds to a triangular section of the plane $\sum_k \lambda_k = 1$. In 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 favours $\alpha_3 > \alpha_2 > \alpha_1$, absolute values increase.

$$Pr(\lambda_1 \ldots \lambda_K) = \frac{\Gamma[\sum_{k=1}^K \alpha_k]}{\prod_{k=1}^K \Gamma[\alpha_k]} \prod_{k=1}^K \lambda_k^{\alpha_k - 1}, \tag{2.9}$$

or for short

$$Pr(\lambda_1 \ldots \lambda_K) = \text{Dir}_{\lambda_{1\ldots K}}[\alpha_1, \alpha_2 \ldots \alpha_K]. \tag{2.10}$$

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.

2.5 Univariate normal distribution

The univariate normal or Gaussian distribution (figure 2.5) 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.

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

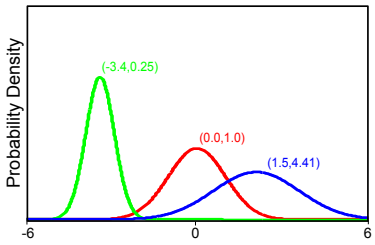

Figure 2.5 The univariate normal distribution is defined on $x \in \mathbb{R}$ and has two parameters (μ, σ^2). The mean parameter μ determines the expected value and the variance σ^2 determines the concentration about the mean so that as σ^2 increases, the distribution becomes wider and flatter.

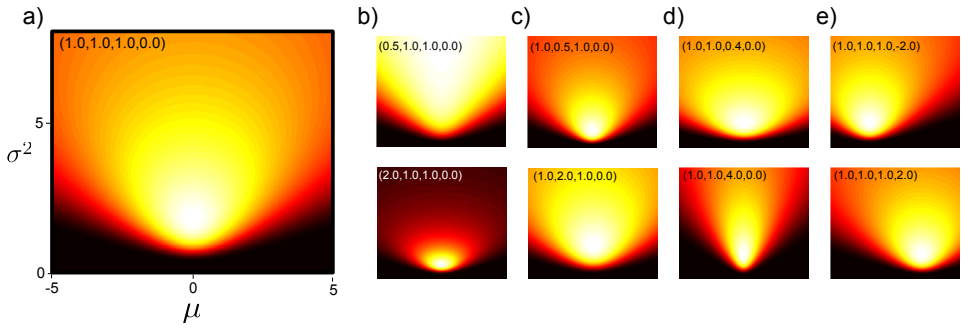

Figure 2.6 The Normal scaled inverse gamma distribution defines a probability distribution over bivariate continuous values μ, σ^2 where $\mu \in [-\infty, \infty]$ and $\sigma^2 \in [0, \infty]$. a) Distribution with parameters $[\alpha, \beta, \gamma, \delta] = [1, 1, 1, 0]$. b) Varying α. c) Varying β. d) Varying γ. e) Varying δ.

parameter σ^2 takes only positive values and determines the width of the distribution. The distribution is defined as

$$Pr(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-0.5(x-\mu)^2/\sigma^2\right] \tag{2.11}$$

and will abbreviate this by writing

$$Pr(x) = \text{Norm}_x[\mu, \sigma^2] \tag{2.12}$$

2.6 Normal inverse gamma distribution

The normal-scaled inverse gamma distribution (figure 2.6) 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.

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. We write

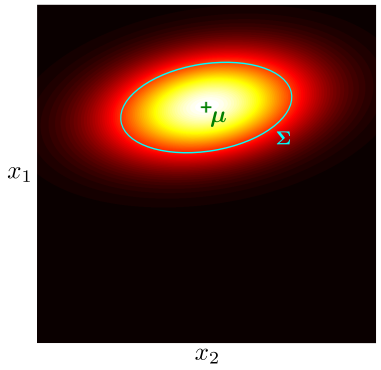

Figure 2.7 The multivariate normal distribution models K-dimensional variables $\mathbf{x} = [x_1 \ldots x_K]^T$ where each dimension x_k is continuous and real. It is defined by a $K \times 1$ vector $\boldsymbol{\mu}$ defining the mean of the distribution and a $K \times K$ covariance matrix $\boldsymbol{\Sigma}$ which determines the shape. The iso-contours of the distribution are ellipsoids where the centre 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.

$$Pr(\mu, \sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta - \mu)^2}{2\sigma^2}\right] \tag{2.13}$$

or for short

$$Pr(\mu, \sigma^2) = \text{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta] \tag{2.14}$$

2.7 Multivariate normal distribution

The multivariate normal or Gaussian distribution models K-dimensional variables $\mathbf{x}$ where each of the K elements $x_1 \ldots x_K$ is continuous and lies in the range $[-\infty, +\infty]$ (figure 2.7). As such the univariate normal distribution is a special case of the multivariate normal where the number of elements K is one. In machine vision the multivariate normal might model the joint distribution of the intensities of K 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 $K \times 1$ vector that describes the mean of the distribution. The covariance $\boldsymbol{\Sigma}$ is a symmetric $K \times K$ 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

$$Pr(\mathbf{x}) = \frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-0.5(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right], \tag{2.15}$$

or for short

$$Pr(\mathbf{x}) = \text{Norm}_{\mathbf{x}}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}\right]. \tag{2.16}$$

The multivariate normal distribution will be used extensively throughout this book, and we devote chapter 4 to describing its properties.

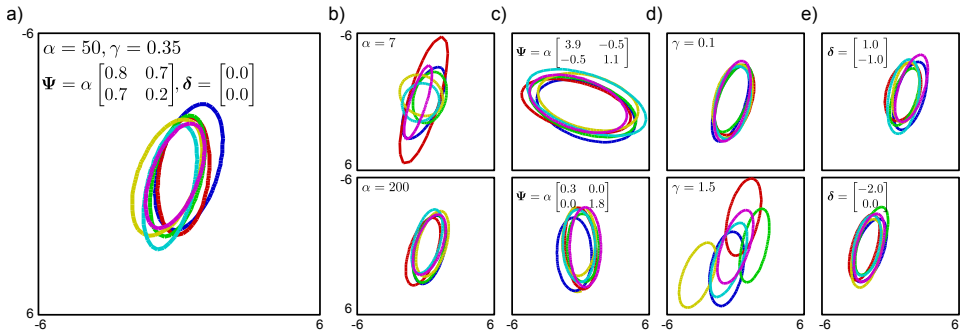

Figure 2.8 Sampling from 2d normal inverse Wishart distribution. a) Each sample consists of a mean vector and covariance matrix, here visualized with 2D ellipses illustrating the isocontour of the associated Gaussian at a Mahalanobis distance of 2. b) Changing α modifies the dispersion of covariances observed. c) Changing $\boldsymbol{\Psi}$ modifies the average covariance. c) Changing γ modifies the dispersion of mean vectors observed. d) Changing $\boldsymbol{\delta}$ modifies the average value of the mean vectors.

2.8 Normal inverse Wishart distribution

The normal inverse Wishart distribution defines a distribution over a $K \times 1$ vector $\boldsymbol{\mu}$ and a $K \times K$ 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 $K \times 1$ vector and $\boldsymbol{\Psi}$ is a $K \times K$ matrix

$$Pr(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Psi^{\alpha/2}|\boldsymbol{\Sigma}|^{-(\alpha+K+2)/2} \exp\left[-0.5\left(2\mathrm{Tr}(\Psi\boldsymbol{\Sigma}^{-1}) - \gamma(\boldsymbol{\mu}-\boldsymbol{\delta})^T\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\delta})\right)\right]}{2^{\alpha K/2}(2\pi)^{K/2}\Gamma_p(\alpha/2)}, \tag{2.17}$$

where Γ_p is the multivariate Gamma function. For short we will write

$$Pr(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{NorIWis}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left[\alpha, \boldsymbol{\Psi}, \gamma, \boldsymbol{\delta}\right]. \tag{2.18}$$

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 2.8).

2.9 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 analagous relationships between the 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 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,

$$\text{Bern}_x[\lambda].\text{Beta}_\lambda[\alpha,\beta] = \kappa(x,\alpha,\beta)\text{Beta}_\lambda\left[\tilde{\alpha},\tilde{\beta}\right] \tag{2.19}$$

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 had the same form. For this case, the relationship in equation 2.19 is easy to prove

$$\begin{aligned}
\text{Bern}_x[\lambda]\text{Beta}_\lambda[\alpha,\beta] &= \lambda^x(1-\lambda)^{1-x}.\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\lambda^{\alpha-1}(1-\lambda)^{\beta-1} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\lambda^{x+\alpha-1}(1-\lambda)^{1-x+\beta-1} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.\frac{\Gamma(x+\alpha)\Gamma(1-x+\beta)}{\Gamma(x+\alpha+1-x+\beta)}\text{Beta}_\lambda[x+\alpha,1-x+\beta] \\
&= \kappa(x,\alpha,\beta)\text{Beta}_\lambda\left[\tilde{\alpha},\tilde{\beta}\right]
\end{aligned} \tag{2.20}$$

The conjugate relationship is important because we do 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 both be computed neatly in closed form.

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. Conjugacy facilitates fitting these distributions to observed data and evaluating new data under the fitted model.

Chapter 3

Fitting probability models

This chapter concerns fitting probability models to observed data $\mathbf{x}_1 \ldots \mathbf{x}_I$ and evaluating the likelihood of a new datum $\mathbf{x}^*$ under the resulting model. This process is referred to as *learning* because we learn about the parameter of the model. We consider three methods: *maximum likelihood*, *maximum a posteriori* and the *Bayesian approach*.

3.1 Maximum likelihood

As the name suggests, the maximum likelihood (ML) method finds the set of parameters $\hat{\boldsymbol{\theta}}$ under which the data $\mathbf{x}_1 \ldots \mathbf{x}_I$ are most likely. To calculate the likelihood $Pr(\mathbf{x}_i|\boldsymbol{\theta})$ for a single data point $\mathbf{x}_i$ we simply evaluate the probability density function at $\mathbf{x}_i$. Assuming each data point was drawn independently, the joint likelihood for a set of points $Pr(\mathbf{x}_{1\ldots I}|\boldsymbol{\theta})$ is the product of the individual likelihoods. Hence, the ML estimate of the parameters is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Pr(\mathbf{x}_{1\ldots I}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{I} \Pr(\mathbf{x}_i|\boldsymbol{\theta}). \tag{3.1}$$

To *evaluate the predictive density* for a new data point $\mathbf{x}^*$ (compute the probability that $\mathbf{x}^*$ belongs to the fitted model), we simply evaluate the probability density function $Pr(\mathbf{x}^*|\hat{\boldsymbol{\theta}})$ using the ML fitted parameters $\hat{\boldsymbol{\theta}}$.

3.2 Maximum a posteriori

In maximum a posteriori (MAP) fitting, we introduce *prior* information about the parameters $\boldsymbol{\theta}$. From previous experience we may know something about the possible parameter values. An obvious example would be for time-sequences: the

values of the parameters at time t tell us a lot about the possible values at time $t+1$ and this information would be encoded in the prior distribution.

As the name suggests, maximum a posteriori estimation maximizes the posterior probability $Pr(\boldsymbol{\theta}|\mathbf{x}_1 \ldots \mathbf{x}_I)$ of the parameters:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} Pr(\boldsymbol{\theta}|\mathbf{x}_1 \ldots \mathbf{x}_I) \\ &= \arg\max_{\boldsymbol{\theta}} \frac{Pr(\mathbf{x}_{1\ldots I}|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_1 \ldots \mathbf{x}_I)} \\ &= \arg\max_{\boldsymbol{\theta}} \frac{\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_1 \ldots \mathbf{x}_I)} \end{aligned} \tag{3.2}$$

where we have used Bayes' rule between the first two lines and subsquently assumed independence of the data likelihoods. In fact, we can discard the denominator as it is constant with respect to the parameters and so does not effect the position of the maximum and we get

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta}). \tag{3.3}$$

Comparing this to the maximum likelihood criterion (equation 3.1) we see that it is identical except for the additional prior term $Pr(\boldsymbol{\theta})$: maximum likelihood is a special case of maximum a posteriori where the prior is uninformative.

The predictive density (likelihood of a new datum $\mathbf{x}^*$ under the fitted model) is again calculated by evaluating the pdf $Pr(\mathbf{x}^*|\hat{\boldsymbol{\theta}})$ using the new parameters.

3.3 The Bayesian approach

In the Bayesian approach we stop trying to estimate fixed, concrete values for the parameters $\boldsymbol{\theta}$ and admit what is obvious: there may be many values of the parameters that are compatible with the data. We compute a probability distribution $Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})$ over the parameters $\boldsymbol{\theta}$ based on data $\mathbf{x}_{1\ldots I}$ using Bayes' rule so that

$$Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I}) = \frac{\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_{1\ldots I})}. \tag{3.4}$$

Evaluating the predictive distribution (the likelihood ascribed to a new datapoint by the model) is more difficult for the Bayesian case: we have not estimated a single model, but found a probability distribution over possible models. Hence, we calculate

$$Pr(x^*|\mathbf{x}_{1\ldots I}) = \int Pr(\mathbf{x}^*|\boldsymbol{\theta})Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})d\boldsymbol{\theta}, \tag{3.5}$$

which can be interpreted as follows: the first term $Pr(\mathbf{x}^*|\boldsymbol{\theta})$ is the prediction for a given value of $\boldsymbol{\theta}$. So, the integral is a weighted sum of the predictions given by different parameters $\boldsymbol{\theta}$, where the weighting is determined by the posterior probability distribution $Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})$ over the parameters.

The predictive density calculations for the Bayesian, MAP and ML cases can be unified if we consider the ML and MAP estimates to be special probability distributions over the parameters where all of the density is at $\hat{\boldsymbol{\theta}}$ (i.e. delta functions at $\hat{\boldsymbol{\theta}}$). The predictive density can now be written:

$$\begin{aligned} Pr(x^*|\mathbf{x}_{1\ldots I}) &= \int Pr(\mathbf{x}^*|\boldsymbol{\theta})\delta(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})d\boldsymbol{\theta} \\ &= Pr(x^*|\hat{\boldsymbol{\theta}}) \end{aligned} \tag{3.6}$$

which is exactly the calculation we originally prescribed.

3.4 Worked example 1: univariate normal

To illustrate the above ideas, we will consider fitting a univariate normal model to scalar data $x_1 \ldots x_I$. Recall that the univariate normal model has pdf

$$Pr(x|\mu,\sigma^2) = \text{Norm}_x[\mu,\sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-0.5\frac{(x-\mu)^2}{\sigma^2}\right], \tag{3.7}$$

and has two parameters, the mean μ and the variance σ^2. Let's generate I independent data points $x_{1\ldots I}$ from a univariate normal with $\mu = 1$ and $\sigma^2 = 1$. Our goal is re-estimate these parameters from the data.

Maximum likelihood estimation

The likelihood $Pr(x_{1\ldots I}|\mu,\sigma^2)$ of the data $x_{1\ldots I}$ for parameters $\{\mu,\sigma^2\}$ is computed by evaluating the pdf for each data point separately and taking the product:

$$\begin{aligned} Pr(x_{1\ldots I}|\mu,\sigma^2) &= \prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2) \\ &= \prod_{i=1}^{I} \text{Norm}_{x_i}[\mu,\sigma^2] \\ &= \frac{1}{(2\pi\sigma^2)^{I/2}}\exp\left[-0.5\sum_{i=1}^{I}\frac{(x_i-\mu)^2}{\sigma^2}\right]. \end{aligned} \tag{3.8}$$

Obviously, the likelihood for some sets of parameters μ,σ^2 will be higher than others (figure 3.1) and it is possible to visualize this by drawing the likelihood as a 2d function of the mean μ and variance σ^2 (figure 3.2). The maximum likelihood solution $\hat{\mu},\hat{\sigma}$ will occur at the peak of this surface so that

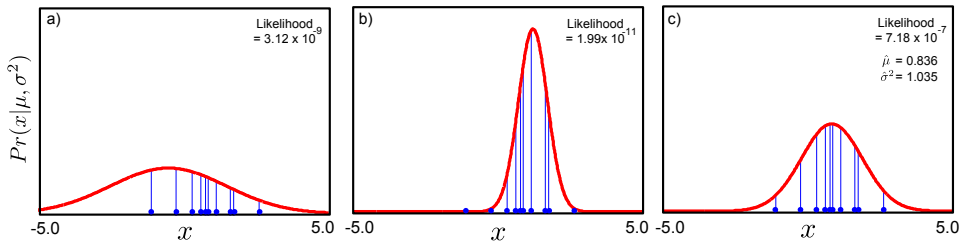

Figure 3.1 Maximum likelihood fitting. The likelihood of a single datapoint is the height of the pdf evaluated at that point (blue vertical lines). The likelihood of a set of independently sampled data is the product of the individual likelihoods. a) The likelihood for this normal distribution is low because the large variance means the height of the pdf is low everywhere. b) The likelihood for this normal distribution is even lower as the left-most datum is very unlikely under the model. c) The maximum likelihood solution is the set of parameters for which the data likelihood is maximized.

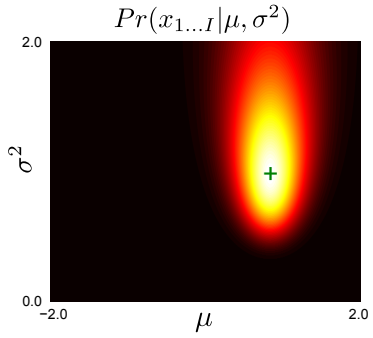

Figure 3.2 The likelihood of the observed data can be plotted as a function of the mean μ and variance σ^2 parameters. The plot shows that there are many parameter settings which might plausibly be responsible for the ten data points from figure 3.1. A sensible choice for the "best" parameter setting is the maximum likelihood solution (green cross) which corresponds to the maximum of this function.

$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu,\sigma^2} Pr(x_{1...I}|\mu, \sigma^2). \tag{3.9}$$

In principle we can maximize this by taking the derivative of equation 3.8 with respect to μ and σ^2, equating the result to zero and solving. In practice however, the resulting equations are messy. To simplify things, we work instead with the logarithm of this expression (the log likelihood, L). Since the logarithm is a monotonic function (figure 3.3), the position of the maximum in the transformed function remains the same. Algebraically, the logarithm turns the product of the likelihoods of the individual data points into a sum and so decouples the contribution of each. The ML parameters can now be calculated as

$$\begin{aligned} \hat{\mu}, \hat{\sigma} &= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{I} \log\left[\text{Norm}_{x_i}[\mu, \sigma^2]\right] \\ &= \arg\max_{\mu,\sigma^2} \left[-0.5I\log[2\pi] - 0.5I\log\sigma^2 - 0.5\sum_{i=1}^{I}\frac{(x_i-\mu)^2}{\sigma^2}\right]. \end{aligned} \tag{3.10}$$

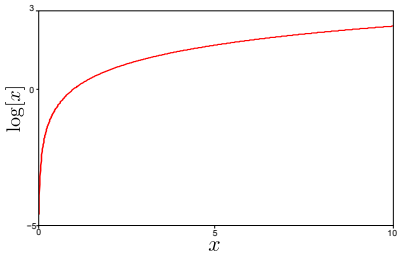

Figure 3.3 The logarithm is a monotonic transformation: if one point is higher than another then it will also be higher after transformation by the logarithmic function. It follows that if we transform the surface in figure 3.2 through the logarithmic function, the maximum will remain in the same position.

To maximize, we differentiate with respect to μ and equate the result to zero

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{I} \frac{(x_i - \mu)}{\sigma^2} = \frac{\sum_{i=1}^{I} x_i}{\sigma^2} - \frac{I\mu}{\sigma^2} = 0 \tag{3.11}$$

and re-arranging, we see that

$$\hat{\mu} = \frac{\sum_{i=1}^{I} x_i}{I}. \tag{3.12}$$

By a similar process the expression for the variance can be shown to be

$$\sigma^2 = \sum_{i=1}^{I} \frac{(x_i - \hat{\mu})^2}{I}. \tag{3.13}$$

These expressions are hardly surprising, but the same idea can be used to estimate parameters in other distributions where the results are less familiar. Figure 3.1 shows a set of data points and three possible fits to the data. The mean of the maximum likelihood fit is the mean of the data. The ML fit is neither too narrow (giving very low probabilities to the furthest data points from the mean) nor too wide (resulting in a flat distribution and giving low probability to all points.

Maximum a posteriori estimation

To find maximum a posteriori parameters we use

$$\begin{aligned} \hat{\mu}, \hat{\sigma}^2 &= \arg\max_{\mu,\sigma^2} \prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2) Pr(\mu,\sigma^2) \\ &= \arg\max_{\mu,\sigma^2} \prod_{i=1}^{I} \text{Norm}_{x_i}[\mu,\sigma^2] \text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta], \end{aligned} \tag{3.14}$$

where we have chosen normal inverse gamma prior with parameters $\alpha, \beta, \gamma, \delta$ (figure 3.4) as this is conjugate to the normal distribution. The expression for the prior is

$$Pr(\mu,\sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta-\mu)^2}{2\sigma^2}\right]. \tag{3.15}$$

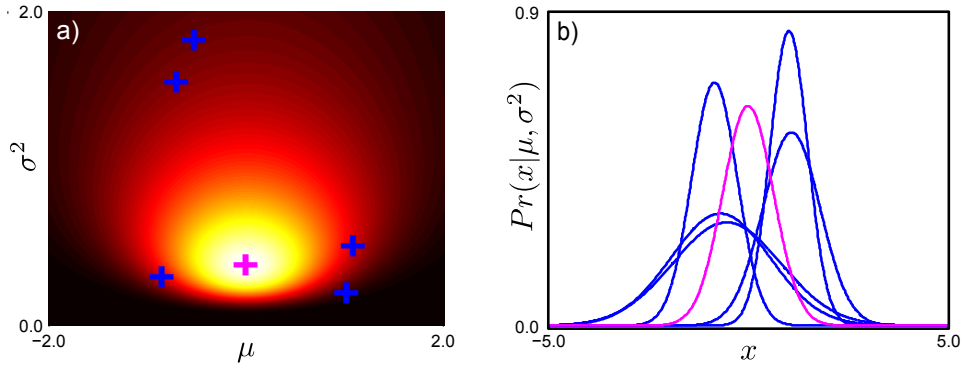

Figure 3.4 a) A normal inverse gamma with $\alpha, \beta, \gamma = 1$ and $\delta = 0$ gives a broad prior distribution over univariate normal parameters. The magenta cross indicates the peak of distribution. The blue crosses are 5 samples randomly drawn from the distribution. b) The samples and peak are visualized as the normal distributions they represent.

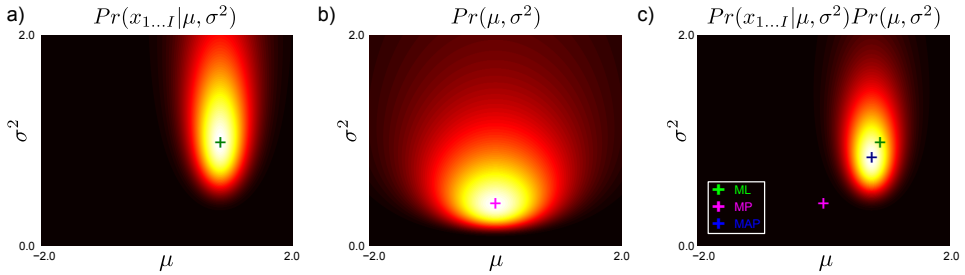

Figure 3.5 a) The likelihood surface is multiplied by b) the prior probability to give a new surface c) that is proportional to the posterior distribution. The maximum a posteriori (MAP) solution (blue cross) is found at the peak of the posterior distribution. It lies between the maximum likelihood (ML) solution (green cross) and the maximum of the prior (purple cross).

The posterior distribution is proportional to the product of the data likelihood and the prior (figure 3.5), and has the highest density in regions that both agree with the data *and* were a priori plausible.

As for the ML case, it is easier to maximize the logarithm of equation 3.14 so that

$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu,\sigma^2} \sum_{i=1}^{I} \log[\text{Norm}_{x_i}\left[\mu, \sigma^2\right]] + \log\left[\text{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]\right]. \tag{3.16}$$

To find the MAP parameters, we substitute in the expressions, differentiate with respect to μ and σ, equate to zero and rearrange to give

$$\hat{\mu} = \frac{\sum_{i=1} x_i + \gamma\delta}{I+\gamma} \qquad \text{and} \qquad \hat{\sigma^2} = \frac{\sum_{i=1}^{I}(x_i-\mu)^2 + 2\beta + \gamma(\delta-\mu)^2}{I+3+2\alpha}. \quad (3.17)$$

The formula for the mean can be more easily understood if we write it as

$$\hat{\mu} = \frac{I\overline{x} + \gamma\delta}{I+\gamma}. \quad (3.18)$$

This is a weighted sum of two terms. The first term is the data mean $\overline{x}$ and is weighted by the number of training examples I. The second term is δ, the value of μ favored by the prior, and is weighted by γ.

This gives some insight into the behaviour of the MAP estimate (figure 3.6). With a large amount of data, the first term dominates and the MAP estimate $\hat{\mu}$ is very close to the data mean (and the ML estimate). With intermediate amounts of data, the $\hat{\mu}$ is a weighted sum of the prediction from the data and the prediction from the prior. With no data at all, the estimate is completely governed by the prior. The hyperparameter γ controls the concentration of the prior with respect to μ and determines the extent of its influence. Similar conclusions can be drawn about the MAP estimate of the variance.

A particularly interesting case occurs where there is a single data point (figure 3.6e-f). The data tells us nothing about the variance and the maximum likelihood estimate $\hat{\sigma}^2$ is zero. This is unrealistic, not least because it accords the datum an infinite likelihood. However, MAP estimation is still valid: $\hat{\sigma}^2$ is determined purely by the prior.

The Bayesian approach

In the Bayesian approach we calculate a posterior distribution $Pr(\mu, \sigma^2|x_{1...I})$ over possible parameter values using Bayes' rule,

$$\begin{aligned} Pr(\mu,\sigma^2|x_{1...I}) &= \frac{\prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2)Pr(\mu,\sigma^2)}{Pr(x_{1...I})} \\ &= \frac{\prod_{i=1}^{I} \text{Norm}_{x_i}[\mu,\sigma^2]\text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]}{Pr(x_{1...I})} \\ &= \frac{\kappa(\alpha,\beta,\gamma,\delta,x_{1...I}).\text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}]}{Pr(x_{1...I})}, \end{aligned} \quad (3.19)$$

where we have used the conjugate relationship between likelihood and prior. The product of the normal likelihood and normal inverse gamma prior creates a posterior over μ, σ^2, which is a new normal inverse gamma distribution, with parameters

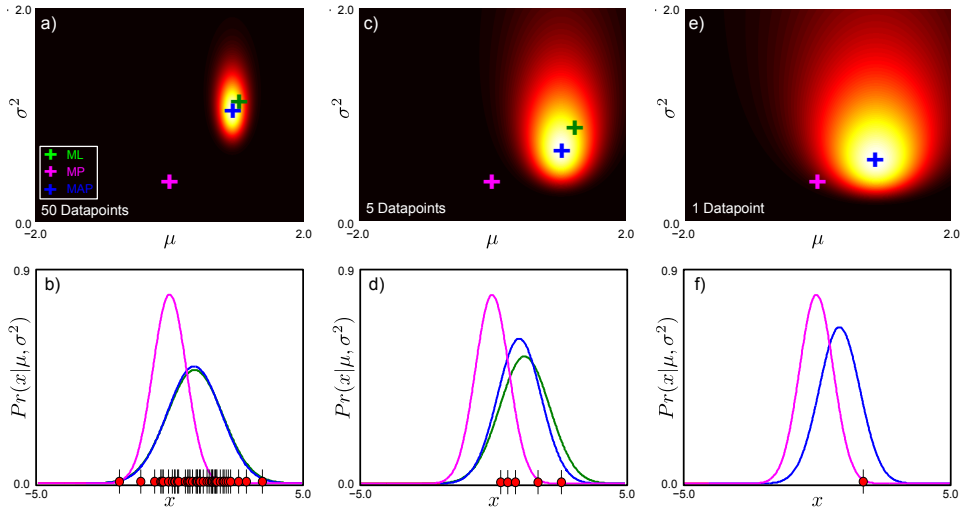

Figure 3.6 Maximum a posteriori estimation. a) MAP solution (blue cross) lies between ML (green cross) and densest region of prior. b) Normal distributions corresponding to MAP solution, ML solution and peak of prior. c-d) With fewer datapoints, the prior has a greater effect on the final solution. e-f) With only one datapoint, the maximum likelihood solution cannot be computed (you cannot calculate the variance of a single point). However, the MAP solution can still be calculated.

$$\tilde{\alpha} = \alpha + I/2, \qquad \tilde{\gamma} = \gamma + I \qquad \tilde{\delta} = \frac{(\gamma\delta + \sum_i x_i)}{\gamma + I}$$
$$\tilde{\beta} = \frac{\sum_i x_i^2}{2} + \beta + \frac{\gamma\delta^2}{2} - \frac{(\gamma\delta + \sum_i x_i)^2}{2(\gamma + I)}. \tag{3.20}$$

Note that the posterior (equation 3.19) must be a valid probability distribution and sum to one, so the constant κ and the denominator must exactly cancel to give

$$Pr(\mu, \sigma^2|x_{1\ldots I}) \quad = \quad \text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}]. \tag{3.21}$$

Now we see the major advantage of using a conjugate prior: we are guaranteed a convenient closed form for the posterior distribution over parameters. This posterior distribution represents the relative plausibility of various parameter settings μ, σ^2 having created the data. At the peak of the distribution is the MAP estimate, but there are many other plausible configurations (figure 3.6).

When data is plentiful, the parameters are well specified and the probability distribution concentrated. In this case, placing all of the probability mass at the MAP estimate is a good approximation to the posterior. However, when data is scarce, many possible parameters might have explained the data and the posterior is broad. In this case approximation with a point mass is inadequate.

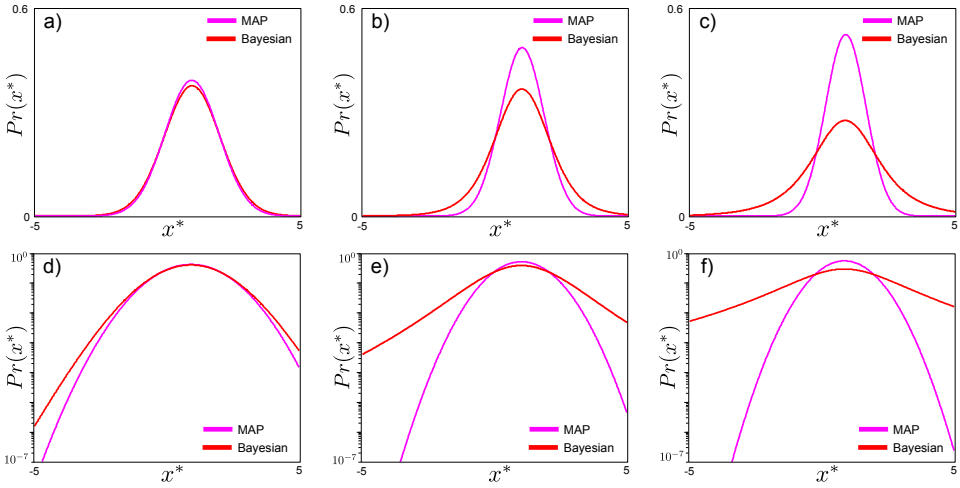

Figure 3.7 a-c) Predictive densities for MAP and Bayesian approaches with 50, 5 and 1 training example. As the training data decreases, the Bayesian prediction becomes less certain but the MAP prediction is erroneously overconfident. d-f) This effect is even more clear on a log scale.

Predictive density

For the maximum likelihood and MAP estimates we evaluate the predictive density (probability that a new data point x^* belongs to the same model) by simply evaluating the pdf of the estimated Gaussian. For the Bayesian case, we compute a weighted average of the predictions for each possible parameter set, where the weighting is given by the posterior distribution over parameters (figure 3.6a-c),

$$\begin{aligned} Pr(x^*|x_{1\ldots I}) &= \iint Pr(x^*|\mu,\sigma^2)Pr(\mu,\sigma^2|x_{1\ldots I})d\mu d\sigma \\ &= \iint \text{Norm}_{x^*}[\mu,\sigma^2]\text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}]d\mu d\sigma \\ &= \iint \kappa(\alpha,\beta,\delta,\gamma,x_{1\ldots I|}).\text{NormInvGam}_{\mu,\sigma^2}[\breve{\alpha},\breve{\beta},\breve{\gamma},\breve{\delta}]d\mu d\sigma. \end{aligned} \tag{3.22}$$

Here we have used the conjugate relation for a second time. The integral contains a constant with respect to μ, σ multiplied by a probability distribution. Taking the constant outside the integral we get

$$\begin{aligned} Pr(x^*|x_{1\ldots I}) &= \kappa(\alpha,\beta,\delta,\gamma,x_{1\ldots I|}) \iint \text{NormInvGam}_{\mu,\sigma^2}[\breve{\alpha},\breve{\beta},\breve{\gamma},\breve{\delta}]d\mu d\sigma \\ &= \kappa(\alpha,\beta,\delta,\gamma,x_{1\ldots I|}) \end{aligned} \tag{3.23}$$

which follows because the integral of a pdf is one. It can be shown that the constant is given by

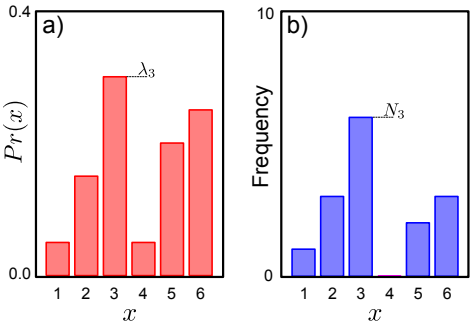

Figure 3.8 a) Categorical probability distribution over 6 discrete values with parameters $\lambda_{1\ldots6}$ where $\sum_{k=1}^{6}\lambda_k = 1$. This could be the relative probability of a biased die landing on its six sides. b) Fifteen observations $x_{1\ldots I}$ randomly sampled from this distribution. We denote the number of times category k is observed by N_k.

$$\kappa(\alpha,\beta,\delta,\gamma,x_{1\ldots I|}) = \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\tilde{\gamma}}\tilde{\beta}^{\tilde{\alpha}}}{\sqrt{\breve{\gamma}}\breve{\beta}^{\breve{\alpha}}}\frac{\Gamma[\breve{\alpha}]}{\Gamma[\tilde{\alpha}]}, \tag{3.24}$$

where

$$\begin{aligned}\breve{\alpha} &= \tilde{\alpha} + 1/2, \qquad & \breve{\gamma} &= \tilde{\gamma} + 1\\ \breve{\beta} &= \frac{x^{*2}}{2} + \tilde{\beta} + \frac{\tilde{\gamma}\tilde{\delta}^2}{2} - \frac{(\tilde{\gamma}\tilde{\delta}+x^*)^2}{2(\tilde{\gamma}+1)}.\end{aligned} \tag{3.25}$$

Figure 3.7 shows the predictive distribution for the Bayesian and MAP cases, for varying amounts of training data. With plenty of training data, there is little difference but as the data decreases, the Bayesian likelihood has a significantly longer tail. This is typical of Bayesian solutions: they are more moderate (less certain) in their predictions. In the MAP case, erroneously committing to a single estimate of μ, σ^2 causes overconfidence in our future predictions.

3.5 Worked example 2: categorical distribution

As a second example, we consider discrete data $x_{1\ldots I}$ where $x_i \in \{1, 2, \ldots, 6\}$ (figure 3.8). This could represent observed rolls of a die with unknown bias. We will describe the data using a categorical distribution (normalized histogram) where

$$Pr(x = k|\lambda_1 \ldots \lambda_K) = \lambda_k. \tag{3.26}$$

For the ML and MAP techniques we estimate the 6 parameters $\lambda_{1\ldots6}$. For Bayesian approach, we compute a probability distribution over the parameters.

Maximum Likelihood

To find the maximum likelihood solution we maximize the product of the individual data likelihoods with respect to the parameters $\lambda_1 \ldots \lambda_k$.

$$\begin{aligned}\hat{\lambda}_1 \ldots \hat{\lambda}_6 &= \arg\max_{\lambda_1 \ldots \lambda_6, \nu} \prod_{i=1}^{I} Pr(x_i|\lambda_1 \ldots \lambda_6) \\ &= \arg\max_{\lambda_1 \ldots \lambda_6, \nu} \prod_{i=1}^{I} \mathrm{Cat}_{x_i}[\lambda_1 \ldots \lambda_6] \\ &= \arg\max_{\lambda_1 \ldots \lambda_6, \nu} \prod_{k=1}^{6} \lambda_k^{N_k} \end{aligned} \tag{3.27}$$

where N_k is the total times we observed bin K in the training data. As before, it is easier to maximize the log probability so that

$$\hat{\lambda}_1 \ldots \hat{\lambda}_6 = \arg\max_{\lambda_1 \ldots \lambda_6} \textstyle\sum_{k=1}^{6} N_k \log[\lambda_k] + \nu(\sum_{k=1}^{6} \lambda_k - 1). \tag{3.28}$$

Note that the parameters must be constrained so that $\sum_{k=1}^{6} \lambda_k = 1$ and we use the Lagrange multiplier ν to enforce this constraint. Taking the derivative with respect to λ_k and ν, equating the resulting equations to zero and re-arranging yields

$$\hat{\lambda}_k = \frac{N_k}{\sum_{k=1}^{6} N_k}. \tag{3.29}$$

In other words, λ_k is the proportion of times that we observed bin k.

Maximum a posteriori

To find the maximum a posteriori solution we need to define a prior. We choose the Dirichlet distribution as it is conjugate to the categorical likelihood. We choose hyperparameters $\alpha_{1\ldots6} = 1$ which gives a uniform prior. This prior over the six categorical parameters is hard to visualize but samples can be drawn and examined (figure 3.9a-e). The MAP solution is given by

$$\begin{aligned}\hat{\lambda}_1 \ldots \hat{\lambda}_6 &= \arg\max_{\lambda_1 \ldots \lambda_6, \nu} \prod_{i=1}^{I} Pr(x_i|\lambda_1 \ldots \lambda_6) Pr(\lambda_1 \ldots \lambda_6) \\ &= \arg\max_{\lambda_1 \ldots \lambda_6, \nu} \prod_{i=1}^{I} \mathrm{Cat}_{x_i}[\lambda_1 \ldots \lambda_6] \mathrm{Dir}_{\lambda_1 \ldots \lambda_6}[\alpha_1 \ldots \alpha_6] \\ &= \arg\max_{\lambda_1 \ldots \lambda_6, \nu} \prod_{k=1}^{6} \lambda_k^{N_k} \prod_{k=1}^{6} \lambda^{\alpha_k - 1} \\ &= \arg\max_{\lambda_1 \ldots \lambda_6, \nu} \prod_{k=1}^{6} \lambda_k^{N_k + \alpha_k - 1} \end{aligned} \tag{3.30}$$

By a similar process to that for the maximum likelihood case, the MAP estimate of the parameters can be shown to be

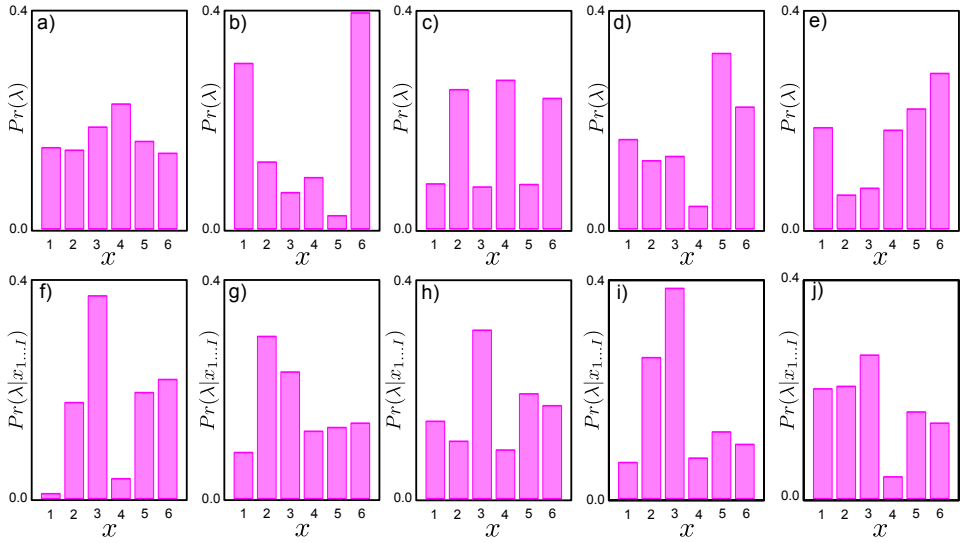

Figure 3.9 a-e) Five samples drawn from Dirichlet prior with hyperparameters $\alpha_{1...6} = 1$. This defines a uniform prior, so each sample looks like a random unstructured probability distribution. f-j) Five samples from Dirichlet posterior. The distribution favours histograms where bin three is larger and bin four is small as suggested by the data.

$$\hat{\lambda}_k = \frac{N_k + \alpha_k - 1}{\sum_{k=1}^{6}(N_k + \alpha_k - 1)} \tag{3.31}$$

where N_k is the number of times that observation k occurred in the training data.

Bayesian Approach

In the Bayesian approach we calculate a posterior over the parameters

$$\begin{aligned} Pr(\lambda_1 \ldots \lambda_6 | x_{1...I}) &= \frac{\prod_{i=1}^{I} Pr(x_i|\lambda_1 \ldots \lambda_6) Pr(\lambda_1 \ldots \lambda_6)}{Pr(x_{1...I})} \\ &= \frac{\prod_{i=1}^{I} \mathrm{Cat}_{x_i}[\lambda_1 \ldots \lambda_6] \mathrm{Dir}_{\lambda_{1...6}}[\alpha_1 \ldots \alpha_6]}{Pr(x_{1...I})} \\ &= \frac{\kappa(\alpha_{1...6}, x_{1...I}) \mathrm{Dir}_{\lambda_{1...6}}[\tilde{\alpha}_{1...6}]}{Pr(x_{1...I})} \\ &= \mathrm{Dir}_{\lambda_{1...6}}[\tilde{\alpha}_{1...6}], \end{aligned} \tag{3.32}$$

where $\tilde{\alpha}_k = N_k + \alpha_k$. We have again exploited the conjugate relationship to yield a posterior distribution with the same form as the prior. The constant κ must again cancel with the denominator to yield a valid probability distribution. Samples from this distribution are shown in figure 3.9f-j).

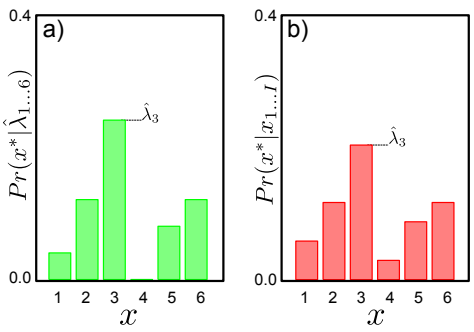

Figure 3.10 Predictive distributions with $\alpha_{1...6} = 1$ for a) Maximum likelihood / maximum a posteriori approaches and b) Bayesian approach. The ML/MAP approach predict the same distribution that exactly follows the data frequencies. The Bayesian approach predicts a more moderate distribution and allots some probability to the case $x = 4$ despite having seen no training examples in this category.

Predictive Density

For the ML and MAP estimates we evaluate the predictive density (probability that a new data point x^* belongs to the same model) by simply evaluating the categorical pdf with the estimated parameters. With the uniform prior ($\alpha_{1...6} = 1$) the MAP and ML predictions are identical (figure 3.10a) and both are exactly proportional to the frequencies of the observed data .

For the Bayesian case, we compute a weighted average of the predictions for each possible parameter set, where the weighting is given by the posterior distribution over parameters so that

$$\begin{aligned} Pr(x^*|x_{1...I}) &= \int Pr(x^*|\lambda_{1...6})Pr(\lambda_{1...6}|x_{1...I})d\lambda_{1...6} \\ &= \int \mathrm{Cat}_{x^*}[\lambda_1 \ldots \lambda_6]\mathrm{Dir}_{\tilde{\lambda}_1 \ldots \tilde{\lambda}_6}[\alpha_{1...6}]d\lambda_{1...6} \\ &= \int \kappa(\alpha_{1...6}, x_{1...I}).\mathrm{Cat}[\breve{\lambda}_{1...k}]d\lambda_{1...K} \\ &= \kappa(\alpha_{1...6}, x_{1...I}). \end{aligned} \tag{3.33}$$

Here, we have again exploited the conjugate relationship to yield a constant multiplied by a probability distribution and the integral is simply the constant as the integral of the pdf is one. For this case, it can be shown that

$$Pr(x^* = k|y = 1) = \kappa(\alpha_{1...6}, x_{1...I}) = \frac{N_k + \alpha_k}{\sum_{j=1}^{20}(N_j + \alpha_j)}. \tag{3.34}$$

This is illustrated in figure 3.10b. It is notable that once more the Bayesian predictive density is less confident than the ML/MAP solutions. In particular, it does not allot zero probability to observing $x^* = 4$ despite the fact that this value was never observed in the training data. This is sensible: just because we have not drawn a four in fifteen observations does not imply that it is inconceivable that we will ever see one: we may have just been unlucky. The Bayesian approach takes this into account and allots this category a small amount of probability.

Summary

We presented three ways to fit a probability distribution to data and to predict the probability of new points. Of the three methods discussed, the Bayesian approach is the most desirable. Here it is not necessary to exactly estimate the parameters and so no errors are introduced because the point estimate is wrong.

However, the Bayesian approach is only tractable when we have a conjugate prior, which makes it easy to calculate the posterior distribution over the parameters $Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})$ and also to evaluate the integral in the predictive density. When this is not the case, we will usually have to rely on maximum a posteriori estimates. Maximum likelihood estimates can be thought of as a special case of maximum a posteriori estimates in which the prior is uninformative.

Chapter 4

Properties of the multivariate normal

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 2 that the multivariate normal distribution has two parameters: the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The mean $\boldsymbol{\mu}$ is a $K \times 1$ vector that describes the mean of the distribution. The covariance $\boldsymbol{\Sigma}$ is a symmetric $K \times K$ 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

$$Pr(\mathbf{x}) = \frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-0.5(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{4.1}$$

or for short

$$Pr(\mathbf{x}) = \text{Norm}_{\mathbf{x}}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}\right]. \tag{4.2}$$

4.1 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

$$\boldsymbol{\Sigma}_{\text{spher}} = \begin{bmatrix}\sigma^2 & 0 \\ 0 & \sigma^2\end{bmatrix} \quad \boldsymbol{\Sigma}_{\text{diag}} = \begin{bmatrix}\sigma_1^2 & 0 \\ 0 & \sigma_2^2\end{bmatrix} \quad \boldsymbol{\Sigma}_{\text{full}} = \begin{bmatrix}\sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_{22}^2\end{bmatrix}. \tag{4.3}$$

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

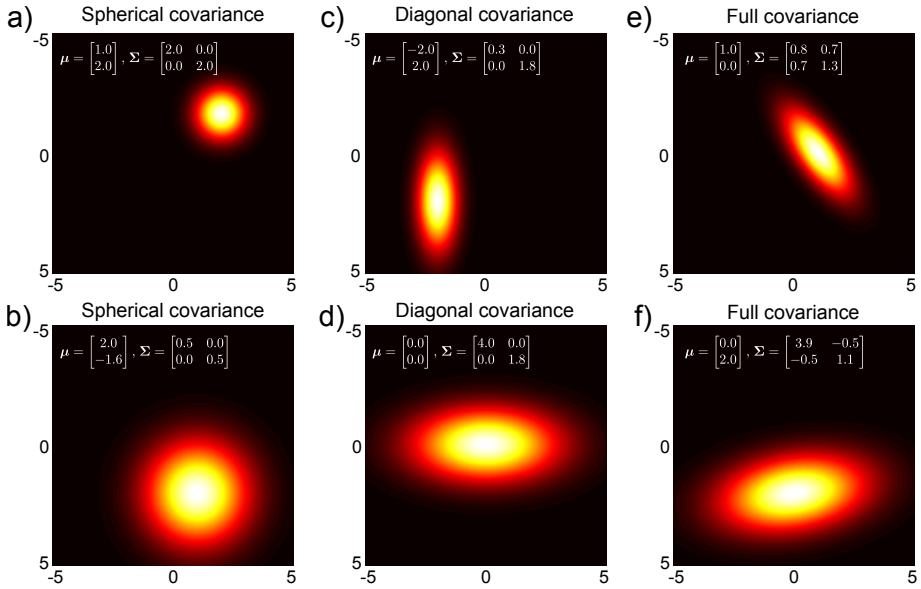

Figure 4.1 Covariance matrices take three forms. a) Spherical covariance matrices are multiples of the identity. The variables are independent and the iso-probability surfaces are hyperspheres (circles in 2d). b) Diagonal covariance matrices permit different non-zero entries on the diagonal, but have 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. c) 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.

value. The full covariance matrix can have non-zero elements everywhere although the matrix is still constrained to be symmetric and positive definite.

For the bivariate case (figure 4.1), spherical covariances produce circular isodensity contours. Diagonal covariances produce ellipsoidal iso-contours are aligned with the coordinate axes. Full covariances also produce ellipsoidal isodensity contours, but these may now take an arbitrary orientation. It is easy to show that the individual variables are independent when the covariance is spherical or diagonal. For example, for the bivariate diagonal case with zero mean we have

$$\begin{aligned} Pr(x_1,x_2) &= \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}}\exp\left[-0.5\begin{pmatrix}x_1 & x_2\end{pmatrix}\boldsymbol{\Sigma}^{-1}\begin{pmatrix}x_1\\x_2\end{pmatrix}\right] \\ &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left[-0.5\begin{pmatrix}x_1 & x_2\end{pmatrix}\begin{pmatrix}\sigma_1^{-2} & 0\\0 & \sigma_2^{-2}\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{x_1^2}{2\sigma_1^2}\right]\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left[-\frac{x_2^2}{2\sigma_2^2}\right] \\ &= Pr(x_1)Pr(x_2) \end{aligned} \tag{4.4}$$

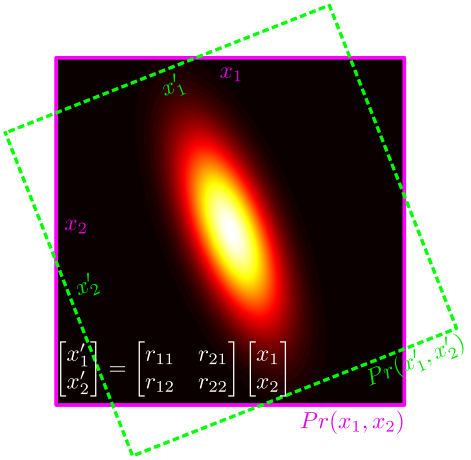

Figure 4.2 Decomposition of full covariance. For every bivariate normal distribution in variables x_1 and x_2 with full covariance matrix, there exists a coordinate system with variables x_1' and x_2' where the covariance is diagonal: the ellipsoidal iso-contours align with the coordinate axes x_1' and 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 $\mathbf{R}\mathbf{\Sigma}'_{diag}\mathbf{R}^T$ of a rotation matrix $\mathbf{R}$ and a diagonal covariance matrix $\mathbf{\Sigma}'_{diag}$.

4.2 Decomposition of covariance

We can use the geometrical intuitions above 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 data from a new set of coordinate axes that *are* aligned with the axes of the normal (figure 4.2): 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

$$Pr(\mathbf{x}') = \frac{1}{(2\pi)^{K/2}|\mathbf{\Sigma}'_{diag}|^{1/2}} \exp\left[-0.5\mathbf{x}'^T\mathbf{\Sigma}'^{-1}_{diag}\mathbf{x}'\right]. \tag{4.5}$$

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

$$\begin{aligned} Pr(\mathbf{x}) &= \frac{1}{(2\pi)^{K/2}|\mathbf{\Sigma}'_{diag}|^{1/2}} \exp\left[-0.5(\mathbf{R}\mathbf{x})^T\mathbf{\Sigma}'^{-1}_{diag}\mathbf{R}\mathbf{x}\right] \\ &= \frac{1}{(2\pi)^{K/2}|\mathbf{R}\mathbf{\Sigma}'_{diag}\mathbf{R}^T|^{1/2}} \exp\left[-0.5\mathbf{x}^T\mathbf{R}^T\mathbf{\Sigma}'^{-1}_{diag}\mathbf{R}\mathbf{x}\right] \end{aligned} \tag{4.6}$$

where we have used $|\mathbf{R}\mathbf{\Sigma}'\mathbf{R}^T| = |\mathbf{R}|.|\mathbf{\Sigma}'|.|\mathbf{R}^T| = 1.|\mathbf{\Sigma}'|.1 = |\mathbf{\Sigma}'|$. Equation 4.6 is a multivariate Gaussian with covariance

$$\mathbf{\Sigma}_{full} = \mathbf{R}\mathbf{\Sigma}'_{diag}\mathbf{R}^T. \tag{4.7}$$

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 $\mathbf{\Sigma}'_{diag}$. Having understood this, it is possible to take an arbitrary valid covariance matrix $\mathbf{\Sigma}_{full}$ and retrieve these elements by calculating the eigenvalue decomposition.

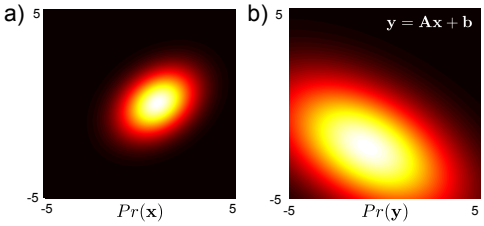

Figure 4.3 Transformation of normal variables. a) If $\mathbf{x}$ has a multivariate normal pdf and we apply a linear transformation to create new variable $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ then b) the distribution of $\mathbf{y}$ is also multivariate normal. The mean and covariance of $\mathbf{y}$ depend on the original mean and covariance of $\mathbf{x}$ and the parameters $\mathbf{A}$ and $\mathbf{b}$

The matrix $\mathbf{R}$ contains the principal directions of the ellipsoid in its rows. The values on the diagonal of $\mathbf{\Sigma}'_{diag}$ encode the variance 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.

4.3 Transformation of variables

The form of the multivariate normal is preserved under linear transformations $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ (figure 4.3). If the original distribution was

$$Pr(\mathbf{x}) = \text{Norm}_{\mathbf{x}}\left[\boldsymbol{\mu}, \mathbf{\Sigma}\right], \tag{4.8}$$

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

$$Pr(\mathbf{y}) = \text{Norm}_{\mathbf{y}}\left[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}^T\mathbf{\Sigma}\mathbf{A}\right]. \tag{4.9}$$

4.4 Marginal distributions

If we marginalize over any subset of random variables in a multivariate normal distribution, the remaining distribution is also normally distributed (figure 4.4). If we partition the original random variable $\mathbf{x}$ into two parts $\mathbf{x}_1$ and $\mathbf{x}_2$ so that

$$Pr(\mathbf{x}) = Pr\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\right) = \text{Norm}_{\mathbf{x}}\left[\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}^T\\ \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{22}\end{bmatrix}\right], \tag{4.10}$$

then

$$\begin{aligned} Pr(\mathbf{x}_1) &= \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11}\right] \\ Pr(\mathbf{x}_2) &= \text{Norm}_{\mathbf{x}_2}\left[\boldsymbol{\mu}_2, \mathbf{\Sigma}_{22}\right]. \end{aligned} \tag{4.11}$$

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.

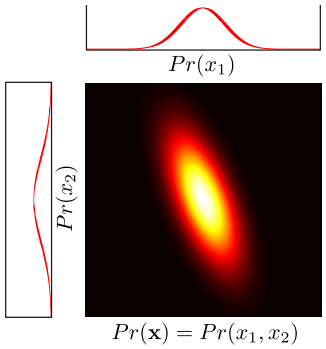

Figure 4.4 The marginal distribution of one subset of variables in a normal distribution is normally distributed. In other words, if we sum over the distribution in any direction, the remaining quantity is also normally distributed. To find the mean and the covariance of the new distribution, we can simply extract the relevant entries from the original mean and covariance matrix.

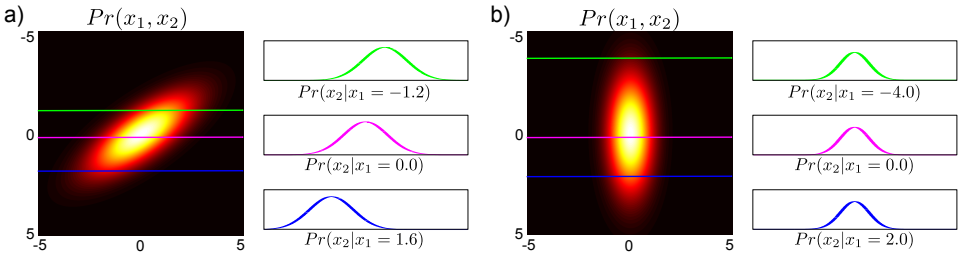

Figure 4.5 Conditional distributions of multivariate normal. a) If we take any multivariate normal distribution, fix a subset of the variables, and look at the distribution of the remaining variables, this distribution will also take the form of a normal. b) If the original multivariate normal has spherical or diagonal variance, the resulting normal distributions are all the same, regardless of the value we conditioned on: these forms of covariance matrix imply independence between the constituent variables.

4.5 Conditional distributions

If the variable $\mathbf{x}$ is distributed as a multivariate normal then the conditional distribution of one subset of variables $\mathbf{x}_1$ with given known values for the remaining variables $\mathbf{x}_2$ is also distributed as a multivariate normal (figure 4.5). If

$$Pr(\mathbf{x}) = Pr\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\right) = \text{Norm}_{\mathbf{x}}\left[\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{21}^T\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right] \tag{4.12}$$

then the conditional distributions are

$$\begin{aligned} Pr(\mathbf{x}_1|\mathbf{x}_2) &= \text{Norm}_{\mathbf{x}_1}\left[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{21}^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}\right] \\ Pr(\mathbf{x}_2|\mathbf{x}_1) &= \text{Norm}_{\mathbf{x}_2}\left[\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}^T\right]. \end{aligned} \tag{4.13}$$

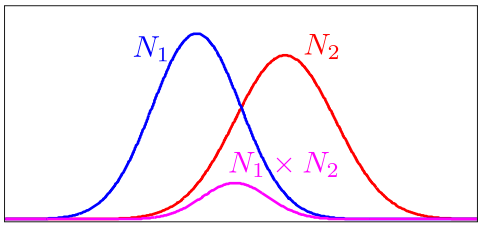

Figure 4.6 The product of any two normals is proportional to a third normal distribution, with a mean between the two original means and a variance that is smaller than either of the original distributions.

4.6 Product of two normals

The product of two normal distributions is proportional to a third normal distribution (figure 4.6). If the two original distributions have means $\mathbf{a}$ and $\mathbf{b}$ and covariances $\mathbf{A}$ and $\mathbf{B}$ respectively then we find that

$$\begin{aligned}\text{Norm}_{\mathbf{x}}[\mathbf{a},\mathbf{A}].\text{Norm}_{\mathbf{x}}[\mathbf{b},\mathbf{B}] &= \\ \kappa.\text{Norm}_x\left[\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}\left(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}\right),\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}\right],\end{aligned} \tag{4.14}$$

where the constant κ can itself be expressed as a normal distribution,

$$\kappa = \text{Norm}_{\mathbf{a}}[\mathbf{b}, \mathbf{A}+\mathbf{B}]. \tag{4.15}$$

4.6.1 Self-conjugacy

The above 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 4.14 that

$$\begin{aligned}\text{Norm}_{\mathbf{x}}[\boldsymbol{\mu},\boldsymbol{\Sigma}].\text{Norm}_{\boldsymbol{\mu}}[\boldsymbol{\mu}_p,\boldsymbol{\Sigma}_p] &= \text{Norm}_{\boldsymbol{\mu}}[\mathbf{x},\boldsymbol{\Sigma}].\text{Norm}_{\boldsymbol{\mu}}[\boldsymbol{\mu}_p,\boldsymbol{\Sigma}_p] \\ &= \kappa.\text{Norm}_{\boldsymbol{\mu}}[\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}]\end{aligned} \tag{4.16}$$

which is the definition of conjugacy (see section 2.9). The new parameters $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are determined from equation 4.14. 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.

4.7 Change of variable

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

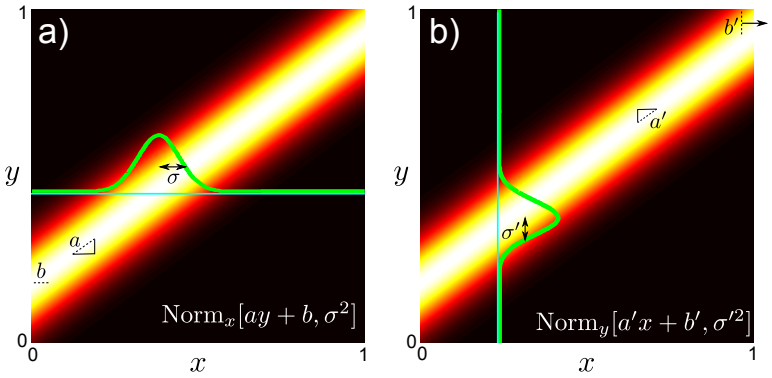

Figure 4.7 a) Consider a normal distribution in x whose variance σ^2 is constant, but whose mean is a linear function $ay + b$ of a second variable y. b) This is mathematically equivalent to a constant κ times a normal distribution in y whose variance σ'^2 is constant and whose mean is a linear function $a'x + b'$ of x.

$$\mathrm{Norm}_{\mathbf{x}}[\mathbf{Ay}+\mathbf{b}, \boldsymbol{\Sigma}] = \kappa.\mathrm{Norm}_{\mathbf{y}}[\mathbf{A}'\mathbf{x}+\mathbf{b}', \boldsymbol{\Sigma}'], \tag{4.17}$$

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

$$\begin{aligned}
\boldsymbol{\Sigma}' &= (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1} && (4.18)\\
\mathbf{A}' &= \boldsymbol{\Sigma}'\mathbf{A}^T\boldsymbol{\Sigma}^{-1} && (4.19)\\
\mathbf{b}' &= -\boldsymbol{\Sigma}'\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}. && (4.20)
\end{aligned}$$

This relationship is mathematically opaque, but is easy to understand visually when x and y are scalars (figure 4.7). It is often used in the context of Bayes' rule where our goal is to move from $Pr(\mathbf{x}|\mathbf{y})$ to $Pr(\mathbf{y}|\mathbf{x})$. It can easily be proved by writing out the terms in the original exponential, extracting quadratic and linear terms in y and completing the square.

Summary

In this chapter we have presented a number of properties of the multivariate normal distribution. These are not the only interesting properties: for example, the Fourier transform of a normal distribution in space is normal in the Fourier domain. The *central limit theorem* states that the sum of a large number of draws from *any* distribution will be normally distributed.

Of the properties discussed in this chapter, two of the most important relate to the marginal and conditional distributions: when we marginalize or take the con-

ditional distribution of a normal with respect to a subset of variables, the resulting distribution is also normal. These properties are exploited in many machine vision applications.

Part II

Machine learning for machine vision

Part II: Machine learning for machine vision

In the following section of this book (chapters 5-8), 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 perverse, 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 5 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 that 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.

We then introduce a simple visual task: we attempt to assign a discrete label to each pixel based on its color using both generative methods (chapter **??**) and discriminative methods (chapter **??**). This label might connote object type (sky, tree, car), material (skin, wood, hair) or the presence of a new object against a known background (foreground, background). The extent to which we can correctly assign such a label based on only the pixel color is obviously quite limited. Nonetheless, this simple task allows demonstrations of the main ideas from Chapter 5.

In chapters 6 and 8 we progress to the more interesting problem of classifying a larger image region. For example, we tackle the problem of face detection. For each square sub-region of an image we decide whether a face is present or absent (assign a label indicating face or non-face).

Chapter 5

A framework for vision tasks

At an abstract level, the goal of computer vision problems is to use the observed image data to infer something about the world. For example, we might observe adjacent frames of a video sequence and wish to infer the camera motion, or we might observe a facial image and wish to infer the identity.

The aim of this chapter is to describe a mathematical framework for solving this type of problem and to organize the resulting models into useful subgroups which will be explored in subsequent chapters.

5.1 The abstract computer vision problem

In vision problems, we take visual data $\mathbf{x}$ and use them to infer the state of the world $\mathbf{y}$. The world state $\mathbf{y}$ may be continuous (the 3d pose of a body model) or discrete (the presence or absence of a particular object). When the state is continuous, we call this *regression*. When the state is discrete, we call this *classification*.

Unfortunately, the measurements $\mathbf{x}$ may be compatible with more than one world state $\mathbf{y}$. The measurement process is noisy and there is inherent ambiguity in visual data: a lump of coal viewed under bright light may produce the same luminance measurements as white paper in dim light. Similarly, a small object seen close-up may produce the same image as a larger object that is further away.

In the face of such ambiguity, the best that we can do is to return the *posterior probability distribution* $Pr(\mathbf{y}|\mathbf{x})$ over possible states $\mathbf{y}$. This describes everything we know about the state after observing the visual data. So, a more precise description of an abstract vision problem is that we wish take observations $\mathbf{x}$ and return the whole posterior probability distribution $Pr(\mathbf{y}|\mathbf{x})$ over world states.

In practice computing the posterior is not always tractable: we often have to settle for returning the world state $\hat{\mathbf{y}}$ at the peak of the posterior (the maximum a posteriori solution). Alternatively, we might draw samples from the posterior: the collection of samples acts as an approximation to the full distribution.

5.1.1 Components of the solution

To solve a vision problem of this kind we need three components.

- We need a *model* that mathematically relates the visual data $\mathbf{x}$ and the world state $\mathbf{y}$. The model specifies a family of possible relationships between $\mathbf{x}$ and $\mathbf{y}$ and the particular relationship is determined by the model parameters $\boldsymbol{\theta}$.
- We need a *learning algorithm* that allows us to fit the parameters $\boldsymbol{\theta}$ using paired training examples $\{\mathbf{x}_i, \mathbf{y}_i\}$ where we know both the measurements and the underlying state.
- We need an *inference algorithm* that takes a new observation $\mathbf{x}$ and uses the model to return the posterior $Pr(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ over the world state $\mathbf{y}$. Alternately, it might return the MAP solution or draw samples from the posterior.

The rest of this book is structured around these components: each chapter focusses on one model and discusses the associated learning and inference algorithms.

5.2 Types of model

The first and most important component of the solution is the model. Every model relating the data $\mathbf{x}$ to the world $\mathbf{y}$ falls into one of three categories. We either:

1. model the contingency of the world on the data $Pr(\mathbf{y}|\mathbf{x})$ or
2. model the joint occurrence of the world and the data $Pr(\mathbf{x}, \mathbf{y})$ or
3. model the contingency of the data on the world $Pr(\mathbf{x}|\mathbf{y})$.

The first type of model is termed *discriminative*. The second two are both termed *generative*: they both maintain probability models over the data which can used to generate (confabulate) new observations. Let's consider these three types of model in turn and discuss learning and inference in each.

5.2.1 Model contingency of world on data (discriminative)

For this case, we first choose an appropriate form for the distribution $Pr(\mathbf{y})$ over the world state $\mathbf{y}$ and then make the distribution parameters a function of the data $\mathbf{x}$. So if the world state was continuous, we might model $Pr(\mathbf{y})$ with a normal distribution and make the mean $\boldsymbol{\mu}$ a function of the data $\mathbf{x}$.

The shape of this function is determined by a second set of parameters, $\boldsymbol{\theta}$. Since the distribution over the state now depends on both the data and these parameters, we write it as $Pr(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ and refer to it as the *posterior distribution*.

The goal of the learning algorithm is to fit the parameters $\boldsymbol{\theta}$ using paired training data $\{\mathbf{x}_i, \mathbf{y}_i\}$. This can be done using maximum likelihood (ML), maximum a posteriori (MAP) or Bayesian approaches (see chapter 3).

The goal of inference is to find a distribution over the possible world states $\mathbf{y}$ for a particular observation $\mathbf{x}$. In this case, this is easy: we have directly constructed an expression for the posterior distribution $Pr(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$.

5.2.2 Model joint occurrence of world and data (generative)

Here we describe the joint probability distribution $Pr(\mathbf{x}, \mathbf{y})$ of the world $\mathbf{y}$ and the data $\mathbf{x}$. For example, if both the world and the state were continuous, we might describe the variable $\mathbf{z} = [\mathbf{x}^T \ \mathbf{y}^T]^T$ with a multivariate normal distribution.

Whatever distribution we choose, it will have some parameters $\boldsymbol{\theta}$. The goal of learning is to use paired training examples $\{\mathbf{x}_i, \mathbf{y}_i\}$ to fit these parameters.

The goal of inference is to compute the posterior distribution $Pr(\mathbf{y}|\mathbf{x})$ and to this end, we use use Bayes' rule

$$Pr(\mathbf{y}|\mathbf{x}) = \frac{Pr(\mathbf{x}, \mathbf{y})}{Pr(\mathbf{x})} = \frac{Pr(\mathbf{x}, \mathbf{y})}{\int Pr(\mathbf{x}, \mathbf{y}) d\mathbf{y}}. \tag{5.1}$$

Algorithms for inference will use this relation to either compute the full posterior, find the MAP world state or draw samples.

5.2.3 Model contingency of data on world (generative)

Here we choose an appropriate form for the distribution $Pr(\mathbf{x})$ over the data $\mathbf{x}$ and make the distribution parameters a function of the world state $\mathbf{y}$. For example, if the data was discrete and multi-valued then we might choose the categorical distribution and make the parameter vector $\boldsymbol{\lambda}$ a function of the world state $\mathbf{y}$.

The shape of this function is determined by a second set of parameters, $\boldsymbol{\theta}$. Since the distribution $Pr(\mathbf{x})$ now depends on both the world state and these parameters, we write it as $Pr(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ and refer to it as the *likelihood*. The goal of learning is to fit the parameters $\boldsymbol{\theta}$ using paired training examples $\{\mathbf{x}_i, \mathbf{y}_i\}$.

In inference, we aim to compute the posterior distribution $Pr(\mathbf{y}|\mathbf{x})$. To this end we specify a prior $Pr(\mathbf{y})$ over the world state (which may itself have parameters $\boldsymbol{\theta}_p$ which need to be learned) and then use Bayes' rule,

$$Pr(\mathbf{y}|\mathbf{x}) = \frac{Pr(\mathbf{x}|\mathbf{y})Pr(\mathbf{y})}{\int Pr(\mathbf{x}|\mathbf{y})Pr(\mathbf{y}) d\mathbf{y}}. \tag{5.2}$$

Algorithms for inference will use this relation to compute the posterior, find the MAP world state or draw samples from the posterior.

Summary

We've seen that there are three distinct approaches to modelling the relationship between the world state $\mathbf{y}$ and the data $\mathbf{x}$, corresponding to modelling the posterior $Pr(\mathbf{y}|\mathbf{x})$, the joint probability $Pr(\mathbf{x}, \mathbf{y})$ or the likelihood $Pr(\mathbf{x}|\mathbf{y})$.

The three model types result in different approaches to inference. For the discriminative model we model the posterior $Pr(\mathbf{y}|\mathbf{x})$ directly and there is no need for further work. For the generative models, we compute the posterior using Bayes' rule. This sometimes results in complex inference algorithms.

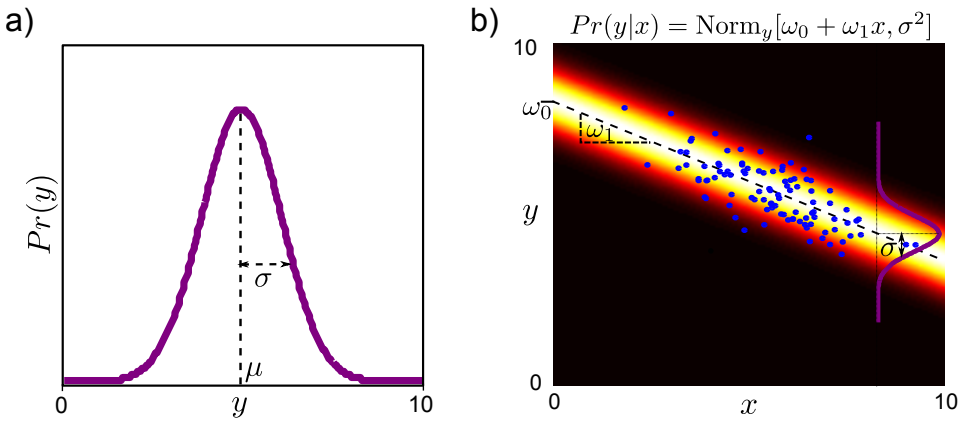

Figure 5.1 Regression by modeling posterior $Pr(y|x)$ (discriminative). a) We represent the world state y with a normal distribution. b) We make the parameters of this distribution a function of the observations x. In this case, the mean is a linear function $\mu = \omega_0 + \omega_1 x$ of the observations and the variance σ^2 is fixed. The associated learning algorithm fits the parameters $\boldsymbol{\theta} = \{\omega_0, \omega_1, \sigma^2\}$ to example training pairs $\{x_i, y_i\}$ (blue dots). In inference we take a new observation x and compute the posterior distribution $Pr(y|x)$ over the state.

The discussion so far has been rather abstract. To clarify things we'll present two example tasks and develop models to solve them using each of the three approaches. In the first example, we'll consider a regression task in which we estimate a univariate continuous state y from a univariate continuous measurement x. In the second example, we'll consider classification: the world state y is now binary and discrete, but we'll assume that the measurement remains univariate and continuous.

5.3 Example 1: Regression

Consider the situation where we make a univariate continuous measurement x from an image and use this to predict a univariate continuous state y. For example, we might want to predict the distance to a car in a road scene based on the total number of pixels in its silhouette.

Model contingency of world on data (discriminative)

We define a probability distribution over the world state y and make its parameters contingent on the data x. Since the world state is univariate and continuous, we'll

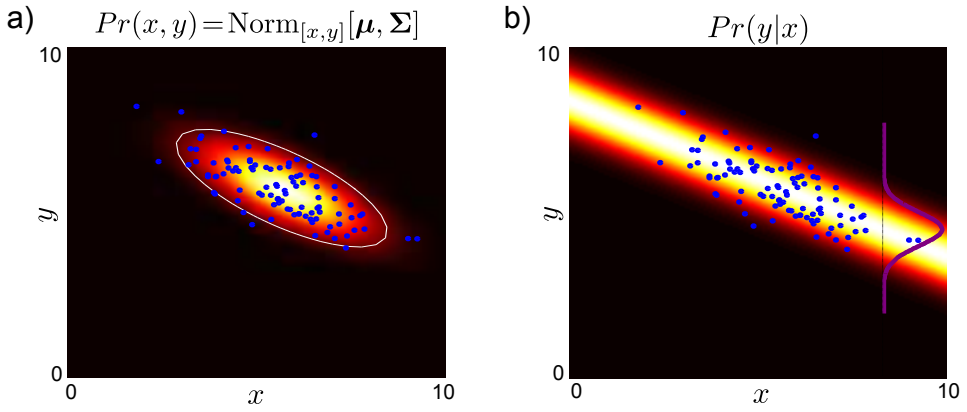

Figure 5.2 Regression by modeling joint distribution $Pr(x, y)$ (generative). a) We model the joint distribution of the measurements x and the world state y with a multivariate normal. The goal of learning is to fit the parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ to paired training examples $\{x_i, y_i\}$ (blue dots). In inference we convert the joint probability to the posterior using Bayes' rule $Pr(y|x) = Pr(x, y)/Pr(x)$. In practice this means normalizing the distribution with respect to x (i.e. normalizing each column).

make the probability distribution a univariate normal. We'll fix the variance, σ^2 and make the mean μ a linear function $\omega_0 + \omega_1 x$ of the data. So we have,

$$Pr(y|x, \boldsymbol{\theta}) = \text{Norm}_y\left[\omega_0 + \omega_1 x, \sigma^2\right] \tag{5.3}$$

were $\boldsymbol{\theta} = \{\omega_0, \omega_1, \sigma^2\}$ are the unknown parameters of the model (figure 5.1).

The associated learning algorithm estimates the model parameters $\boldsymbol{\theta}$ from paired training examples $\{\mathbf{x}_i, \mathbf{y}_i\}$. For example, in the MAP approach we seek

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Pr(\boldsymbol{\theta}|y_{1\ldots I}, x_{1\ldots I}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{I} Pr(y_i|x_i, \boldsymbol{\theta}) Pr(\boldsymbol{\theta}), \tag{5.4}$$

where we have assumed that the I training pairs $\{x_i, y_i\}$ are independent, and defined a suitable prior $Pr(\boldsymbol{\theta})$.

We also need an *inference algorithm* that takes visual data x and returns the posterior distribution $Pr(y|x, \boldsymbol{\theta})$. Here this is very simple: we simply evaluate equation 5.11 using the data x and the learnt parameters $\hat{\boldsymbol{\theta}}$.

Model the joint occurrence of world and data (generative)

We concatenate the measurements and state to form a new variable $\mathbf{z} = [x, y]^T$ and model this with the bivariate normal distribution (figure 5.2)

$$Pr(x, y|\boldsymbol{\theta}) = Pr(\mathbf{z}) = \text{Norm}_{\mathbf{z}}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}\right], \tag{5.5}$$

The associated learning algorithm, fits the parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ of the joint distribution using either the ML, MAP or Bayesian approaches (see chapter 3).

The *inference algorithm* takes a new example of visual data x and returns the posterior distribution $Pr(y|x, \boldsymbol{\theta})$. The posterior is computed from the joint distribution using Bayes' rule $Pr(y|x) = Pr(x, y)/Pr(x)$. For the multivariate normal distribution it is possible to find this in closed form (see section 4.5). It takes the form of a normal distribution with constant variance but a mean that is proportional to the observed data x.

Model the contingency of data on world (generative)

Finally, consider choosing a probability distribution over the data x and making its parameters contingent on the world state y. Since the data is univariate and continuous, we'll model the data as a normal distribution with fixed variance, σ^2 and a mean μ that is linear function $\omega_0 + \omega_1 y$ of the world state (figure 5.3) so that

$$Pr(x|y, \boldsymbol{\theta}) = \text{Norm}_x\left[\omega_0 + \omega_1 y, \sigma^2\right]. \tag{5.6}$$

We also need a prior $Pr(y)$ over the world states which might also be normal so

$$Pr(y) = \text{Norm}_y[\mu_p, \sigma_p^2]. \tag{5.7}$$

The learning algorithm fit the parameters $\boldsymbol{\theta} = \{\omega_0, \omega_1, \sigma^2\}$ using paired training data $\{x_i, y_i\}$ and the parameters $\boldsymbol{\theta}_p = \{\mu_p, \sigma_p^2\}$ using the world states y_i. The inference algorithm takes a new datum x and returns the posterior $Pr(y|x, \boldsymbol{\theta})$ over the world state y using Bayes rule

$$Pr(y|x) = \frac{Pr(y|x)Pr(y)}{Pr(x)} = \frac{Pr(x, y)Pr(y)}{Pr(x)}. \tag{5.8}$$

In this case the posterior can be computed in closed form and is again normally distributed with fixed variance a mean that is proportional to the data x.

Discussion

We have presented three models that can be used to estimate the world state y from an observed data example x, based on modelling the posterior $Pr(y|x)$, the joint probability $Pr(x, y)$ and the likelihood $Pr(x|y)$ respectively.

The three models were carefully chosen so that they predict exactly the same posterior $P(y|x)$ over the world state (compare figures 5.1b, 5.2b and 5.3e). This is only the case with maximum likelihood learning: the MAP approach we would have placed priors on the parameters, and because each model is parameterized differently they would probably have have a different effect in each case.

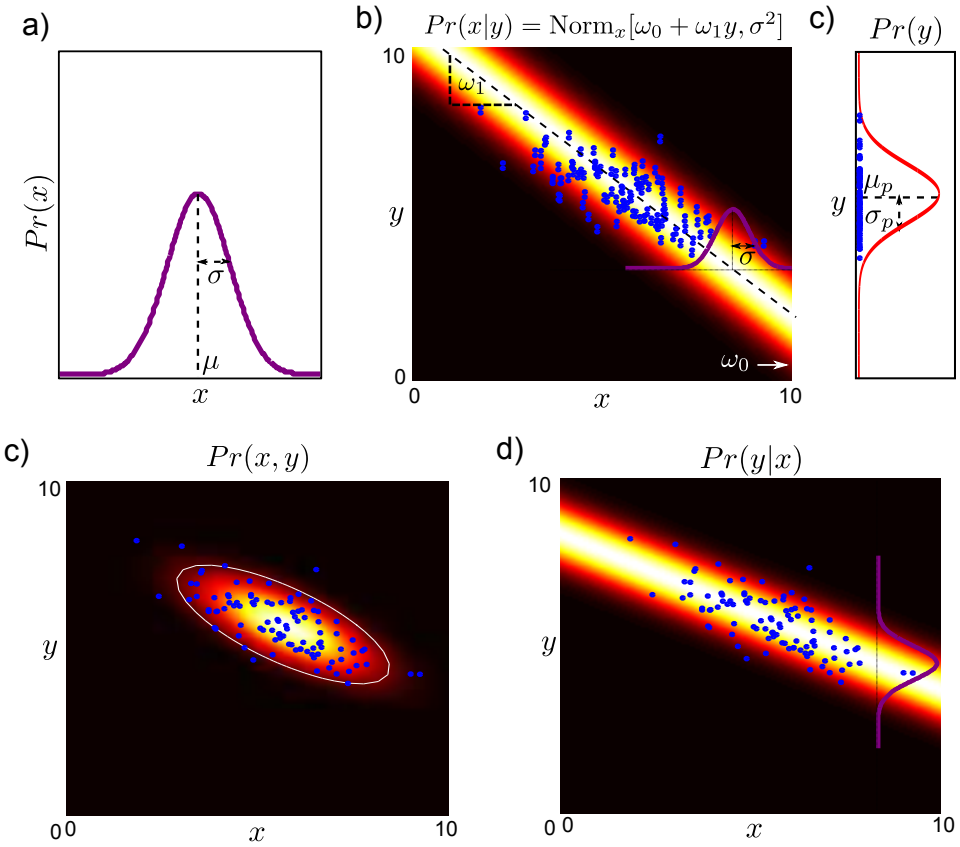

Figure 5.3 Regression by modeling likelihood $Pr(x|y)$ (generative). a) We choose a normal distribution to represent the data x. b) We make the parameters of this distribution depend on the world state y. Here the mean is a linear function $\mu = \omega_0 + \omega_1 y$ of the observations and the variance σ^2 is fixed. The associated algorithm fits the parameters $\boldsymbol{\theta} = \{\omega_0, \omega_1, \sigma^2\}$ to example training pairs $\{x_i, y_i\}$ (blue dots). c) We also learn a prior distribution over the world state y (here modelled as a normal distribution with parameters $\boldsymbol{\theta}_p = \{\mu_p, \sigma_p\}$). In inference we take a new datum x and compute the posterior distribution $Pr(y|x)$ over the state. This can be done by either by (d) computing the joint distribution $Pr(x, y) = Pr(x|y)Pr(y)$ (by weighting each row of (b) by the appropriate value from the prior) and proceeding as for the previous model or (e) using Bayes' rule $Pr(y|x) = Pr(x|y)Pr(y)/Pr(x)$ to directly compute the posterior (weighting each row by the prior and normalizing the columns).

5.4 Example 2: Binary Classification

As a second example, we'll consider the case where the observed measurement x is univariate and continuous, but the world state y is discrete and can take one of

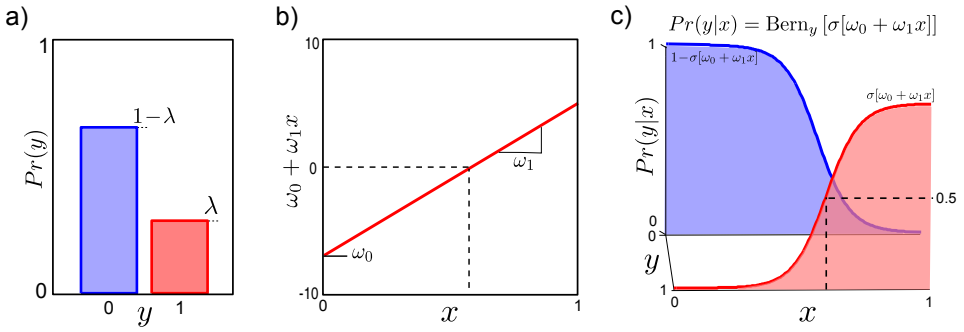

Figure 5.4 Classification by modeling posterior $Pr(y|x)$ (discriminative). a) We represent the world state y as a Bernoulli distribution. We make the Bernoulli parameter λ a function of the observations x. b) To this end we form a linear function $\omega_0 + \omega_1 x$ of the observations. c) The Bernoulli parameter λ is formed by passing the linear function through the logistic sigmoid $\sigma[\bullet]$ to constrain the value to lie between 0 and 1, giving the characteristic sigmoid shape. In learning we fit parameters $\boldsymbol{\theta} = \{\omega_0, \omega_1\}$ using example training pairs $\{x_i, y_i\}$. In inference we take a new datum x and evaluate the posterior $Pr(y|x)$ over the state.

two values. For example, we might wish to classify a pixel as belonging to a skin or non-skin region based on observing just the red channel.

Model contingency of world on data (discriminative)

We define a probability distribution over the world state $y \in \{0, 1\}$ and make its parameters contingent on the data x. Since the world state is discrete and binary, we'll use a Bernoulli distribution. This has a single parameter λ which determines the probability of success (i.e. $Pr(y = 1) = \lambda$).

We make λ a function of the data x, but in doing so we must ensure the constraint $0 \leq \lambda \leq 1$ is obeyed. To this end, we form linear function $\omega_0 + \omega_1 x$ of the data x (which returns a value in the range $[-\infty \;\; \infty]$) and pass this through a function $\sigma()$ that maps $[-\infty \;\; \infty]$ to $[0 \;\; 1]$, so that

$$Pr(y|x) = \text{Bern}_y\left[\sigma[\omega_0 + \omega_1 x]\right] = \text{Bern}_y\left[\frac{1}{1 + \exp[-\omega_0 - \omega_1 x]}\right]. \tag{5.9}$$

The result of this is to produce a sigmoidal dependence of the distribution parameter λ on the data x (see figure 5.4). The function $\sigma[]$ is called the *logistic sigmoid*. The whole model is rather confusingly termed *logistic regression* despite being used here for classification.

In learning, we aim to fit the parameters $\boldsymbol{\theta} = \{\omega_0, \omega_1\}$ from paired training examples $\{x_i, y_i\}$. In inference, we simply substitute in the observed data value x into equation 5.9 to retrieve the posterior distribution $Pr(y|x)$ over the state.

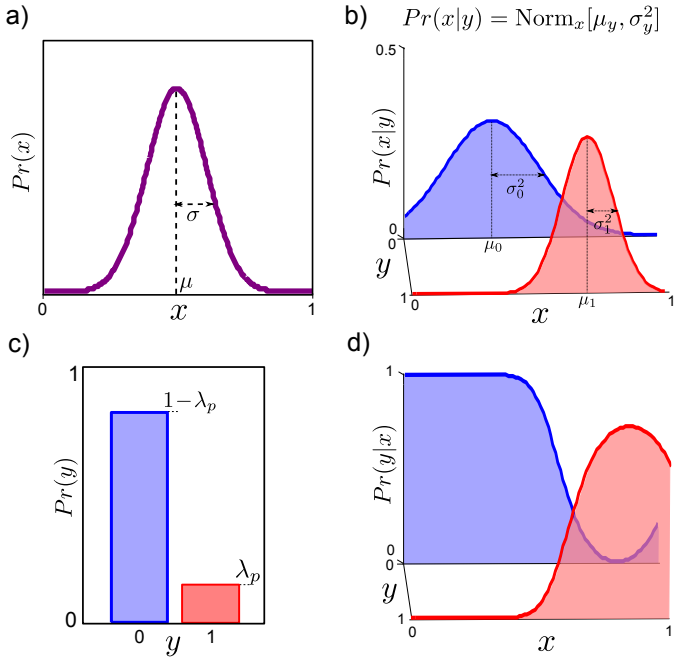

Figure 5.5 Classification by modeling likelihood $Pr(x|y)$ (generative). a) We choose a normal distribution to represent the data x. b) We then make the parameters $\{\mu, \sigma^2\}$ of this distribution a function of the world state y. In practice, this means using one mean and variance when the world state $y = 0$ and another when $y = 1$. The associated learning algorithm fits the parameters $\boldsymbol{\theta} = \{\mu_0, \mu_1, \sigma_0^2, \sigma_1^2\}$ to example training pairs $\{x_i, y_i\}$. We also model the prior probability of the world state y with a Bernoulli distribution with parameter λ_p. In inference we take a new observation x and compute the posterior distribution $Pr(y|x)$ over the state using Bayes' rule.

Model joint occurrence of world and data (generative)

Here our goal is to fit a distribution to the compound variable $\mathbf{z} = [x \;\; y]$. Although this is possible in theory, there are no common probability distributions defined over a concatenation of discrete and continuous variables and this approach is rarely used in practice.

Model contingency of data on world (generative)

We choose a probability distribution over the data x and make its parameters contingent on the world state y. Since the data is univariate and continuous, we'll choose a univariate normal and allow variance, σ^2 and the mean μ to be functions of the binary world state y (figure 5.5) so that the likelihood is

$$Pr(x|y,\boldsymbol{\theta}) = \text{Norm}_x\left[\mu_y, \sigma_y^2\right]. \tag{5.10}$$

In practice this means that we have one set of parameters μ_0, σ_0^2 when the state of the world is $y = 0$ and a different set μ_1, σ_1^2 when the state of the world is $y = 1$ so we can write

$$\begin{aligned} Pr(x|y=0) &= \text{Norm}_x\left[\mu_0, \sigma_0^2\right] \\ Pr(x|y=1) &= \text{Norm}_x\left[\mu_1, \sigma_1^2\right]. \end{aligned} \tag{5.11}$$

These are referred to as *class conditional density functions* as they model the density of the data for each class separately.

We also need a prior distribution $Pr(y)$ over the world states,

$$Pr(y) = \text{Bern}_y[\lambda_p], \tag{5.12}$$

where λ_p is the prior probability of observing state $y = 1$.

In learning we fit the parameters $\boldsymbol{\theta} = \{\mu_0, \sigma_0^2, \mu_1, \sigma_1^2\}$ and using paired training data $\{x_i, y_i\}$. In practice this consists of fitting the class conditional density functions $Pr(x|y = 0)$ from just the data x where the state y was 0, and $P(x|y = 1)$ from the data x where the state was 1. We also learn the parameter λ_p of the prior from the world states y_i.

The associated inference algorithm takes new datum $\mathbf{x}$ and returns the posterior distribution $Pr(y|x,\boldsymbol{\theta})$ over the world state $\mathbf{y}$ using Bayes rule,

$$Pr(y|x) = \frac{Pr(x|y)Pr(y)}{\sum_{y=0}^{1} Pr(x|y)Pr(y)}. \tag{5.13}$$

This is very easy to compute: it consists of evaluating the two class conditional density functions, weighting each by the appropriate prior and normalizing so that these two values sum to one.

Discussion

For binary classification, there is an asymmetry between world state which is discrete and the measurements which are continuous. As a consequence of this, the resulting models look quite different, and the posteriors over the world state y as a function of the data x have different forms (compare figure 5.4c with figure 5.5d). For the discriminative model this function is by definition sigmoidal, but for the generative case it has a more complex form that was implicitly defined by the normal likelihoods. In general, choosing between describing $Pr(y|x)$, $Pr(x, y)$ and $P(x|y)$ will effect the expressiveness of the final model.

5.5 Which type of model should we use?

We have established that there are three different types of model that relate the world state and the observed data and provided two concrete examples in which

we investigated each in turn. So, when should we use each type of model? There is no definitive answer to this question, but some considerations are:

- *Generative methods* build probability models $Pr(\mathbf{x}, \mathbf{y})$ or $P(\mathbf{x}|\mathbf{y})$ over the data whereas *discriminative models* just build a probability model $Pr(\mathbf{y}|\mathbf{x})$ over the world state. The data (usually an image) is generally of much higher dimension than the world state (some aspect of a scene), and modelling it is costly. Moreover, there may be many features of the probability distribution over the data, which do not influence the state: we might devote parameters to describing whether data configuration 1 is more likely than data configuration 2 although that they both imply the same world state.

- Inference is simpler with *discriminative* models as they directly construct a the posterior probability distribution as a function of the data. However, it may still be hard (or even intractable) to find the MAP world state. In contrast generative models compute the posterior indirectly via Bayes' rule and this may result in complex algorithms.

- Modeling the likelihood $Pr(\mathbf{x}|\mathbf{y})$ mirrors the actual way that the data was created: the state of the world did create the observed data through some physical process (usually light being emitted from a source, interacting with the object and being captured by a camera). If we wish to build information about the generation process into our model this approach is desirable. For example, we can account for simple effects such as perspective projection and occlusion. Using the other approaches, it is harder to exploit this knowledge: essentially we have to re-learn these phenomena from the data.

- In some situations, some parts of the training or test data vector $\mathbf{x}$. Here, generative models are preferred. They model the joint distribution over all of the data dimensions and can effectively interpolate the missing dimensions.

- In principle, we can model the same relationship using each of the three approaches but some models are naturally easier to build in one form or the other. For instance, the functional forms relating the distribution over the world to the data in figures 5.4c and 5.5d are different. There is no simple discriminative model that can produce a result like that in figure 5.5d and no simple generative model that produces a result like that in figure 5.4c.

Summary

In this chapter, we have provided an overview of how abstract vision problems are will be solved using machine learning techniques. We have illustrated these ideas with some simple examples. We did not provide the implementation level details of the learning and inference algorithms, these are in a sense just details (they are mostly tackled in subsequent chapters anyway). We hope that the reader would now feel confident to devise a new model to infer the state of the world

from the measurements whatever form these take (discrete, continuous, univariate, multivariate etc.).

In the next few chapters we elaborate on these models. In the following chapter, we investigate building complex probability density models. These are needed for generative models: they can describe both (i) the joint occurrence $Pr(\mathbf{x}, \mathbf{y})$ of the data and world in regression tasks when both are continuous and (ii) the likelihood in classification tasks via the class conditional density functions $Pr(\mathbf{x}|y = k)$. In chapter 7 we investigate discriminative approaches to regression. Finally, in chapter 8 we investigate discriminative approaches to classification.