6

Discrete and Continuous Variables

Overview

Probabilistic models usually include both discrete and continuous quantities. In Section 6.1 we explain how to jointly model the behavior of such quantities by representing them as random variables belonging to the same probability space. Section 6.2 defines the conditional distribution of continuous random variables given the values of discrete variables, and introduces mixture models. Conversely, Section 6.3 explains how to characterize the conditional distribution of discrete random variables given the values of continuous random variables. Section 6.4 discusses independence and conditional independence between discrete and continuous variables. Section 6.5 and 6.6 describe Gaussian discriminant analysis and Gaussian mixture models, two popular probabilistic models for classification and clustering. Section 6.7 introduces the Bayesian framework for parametric modeling.

6.1 Joint Distribution Of Discrete And Continuous Variables

In Section 4.1.1 we explain how to model the joint behavior of multiple discrete uncertain quantities by representing them as discrete random variables in a common probability space. Section 5.1 applies the same approach to continuous quantities, represented by continuous random variables. Here, we show that we can also leverage it to build probabilistic models with both discrete and continuous variables.

Let $\tilde{d} : \Omega \to R_{\tilde{d}}$ be a discrete random variable ($R_{\tilde{d}}$ is the discrete range of $\tilde{d}$) and $\tilde{c} : \Omega \to \mathbb{R}$ a continuous random variable associated to the same probability space $\{\Omega, \mathcal{C}, \mathrm{P}\}$. Each outcome ω in Ω is mapped to a value $\tilde{d}(\omega) \in R_{\tilde{d}}$ by $\tilde{d}$, and to a real value $\tilde{c}(\omega)$ by $\tilde{c}$. In order to characterize the joint behavior of the random variables, we consider the events that $\tilde{d}$ equals a value d in its discrete range $R_{\tilde{d}}$ and $\tilde{c}$ is in a Borel set (i.e. an interval or union of intervals) C *at the same time*. The event that both $\tilde{d} = d$ and $\tilde{c} \in C$ is the intersection of the events

$$D := \left\{\omega : \tilde{d}(\omega) = d\right\} \qquad \text{and} \qquad C^{-1} := \left\{\omega : \tilde{c}(\omega) \in C\right\}. \tag{6.1}$$

These events belong to the collection $\mathcal{C}$ of the common probability space as long as $\tilde{d}$ is a valid discrete random variable, and $\tilde{c}$ is a valid continuous random variable.

The probability $\mathrm{P}(D \cap C^{-1})$, which we denote by $\mathrm{P}(\tilde{d} = d, \tilde{c} \in C)$, is therefore well defined. This is illustrated in Figure 6.1.

We usually describe the behavior of multiple discrete random variables using their joint pmf (see Section 4.1.2), and the behavior of multiple continuous random variables using their joint pdf (see Section 5.3). The challenge when we consider models that include both discrete and continuous quantities is that neither of these objects is well defined.* The joint cdf is well defined, but unfortunately it is also very annoying to manipulate (see Example 5.5). Instead, we define and work with marginal and conditional distributions, which allow us to characterize the joint distribution using the corresponding marginal and conditional pmfs or pdfs, as explained in the rest of this section.

6.2 Conditional Distribution Of Continuous Variables

Consider a continuous random variable $\tilde{c} : \Omega \to \mathbb{R}$ and a discrete random variable $\tilde{d} : \Omega \to R_{\tilde{d}}$ defined on the same probability space $\{\Omega, \mathcal{C}, \mathrm{P}\}$. As explained in Section 6.1, the events $\{\tilde{c} \leq c\}$ and $\{\tilde{d} = d\}$, where $c \in \mathbb{R}$ and $d \in R_{\tilde{d}}$ are well defined. If $\mathrm{P}(\{\tilde{d} = d\}) \neq 0$, then the conditional probability of $\{\tilde{c} \leq c\}$ given $\{\tilde{d} = d\}$ is also well defined. This allows us to define the conditional cdf of $\tilde{c}$ given $\tilde{d} = d$. If the conditional cdf is differentiable, then its derivative can be interpreted as a conditional probability density, equal to the ratio between the conditional probability that $\tilde{c}$ belongs to a small interval of length ϵ given $\tilde{d} = d$ and the length of the interval (see Section 3.3),

$$\frac{\mathrm{d}F_{\tilde{c}\,|\,\tilde{d}}\,(c\,|\,d)}{\mathrm{d}c} := \lim_{\epsilon \to 0} \frac{F_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d) - F_{\tilde{c}\,|\,\tilde{d}}(c - \epsilon\,|\,d)}{\epsilon} \tag{6.2}$$

$$= \lim_{\epsilon \to 0} \frac{\mathrm{P}(c - \epsilon < \tilde{c} \leq c\,|\,\tilde{d} = d)}{\epsilon}. \tag{6.3}$$

Definition 6.1 (Conditional cdf and pdf of a continuous random variable given a discrete random variable)**.** *Let $\tilde{c}$ and $\tilde{d}$ be a continuous and a discrete random variable defined on the same probability space. The conditional cdf and pdf of $\tilde{c}$, given $\tilde{d}$ are*

$$F_{\tilde{c}\,|\,\tilde{d}}\,(c\,|\,d) := \mathrm{P}\left(\tilde{c} \leq c\,|\,\tilde{d} = d\right), \tag{6.4}$$

$$f_{\tilde{c}\,|\,\tilde{d}}\,(c\,|\,d) := \frac{\mathrm{d}F_{\tilde{c}\,|\,\tilde{d}}\,(c\,|\,d)}{\mathrm{d}c}, \tag{6.5}$$

for d such that $p_{\tilde{d}}(d) \neq 0$, assuming $F_{\tilde{c}\,|\,\tilde{d}}$ is differentiable.

Figure 6.2 shows the conditional pdf of the temperature on the Mauna Loa volcano in Hawaii depending on whether there is precipitation or not, estimated using kernel density estimation (see Section 3.5.2) using data extracted from Dataset 9. The two conditional pdfs look very different.

*It is possible to define a probability density for discrete random variables using Dirac deltas, but this requires leveraging the machinery of distributions, and does not provide any advantage within the scope of this book.

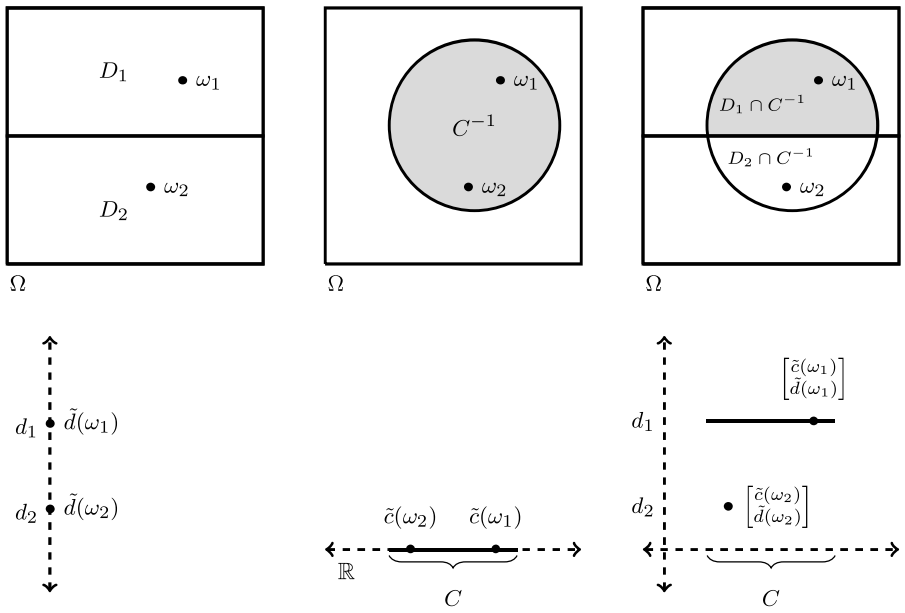

Figure 6.1 Joint distribution of discrete and continuous random variables. The discrete random variable $\tilde{d}$ and the continuous random variable $\tilde{c}$ map outcomes in the sample space Ω to a discrete set $\mathbb{R}_{\tilde{d}} := \{d_1, d_2\}$ and the real line $\mathbb{R}$, respectively. The top left Venn diagram shows the partition $\{D_1, D_2\}$ of Ω associated to $\tilde{d}$. The outcomes $\omega_1 \in D_1$ and $\omega_2 \in D_2$ are mapped to $\tilde{d}(\omega_1) = d_1$ and $\tilde{d}(\omega_2) = d_2$ respectively. The Venn diagram in the top middle image shows the event of outcomes C^{-1} mapped by $\tilde{c}$ to the Borel set C, depicted below on the horizontal real line. Both ω_1 and ω_2 are mapped to C by $\tilde{c}$, so they belong to C^{-1}. The Venn diagram on the right shows that the outcomes mapped to d_1 by $\tilde{d}$ and simultaneously to C by $\tilde{c}$ are in the intersection $D_1 \cap C^{-1}$, which includes ω_1. In $\mathbb{R}^2$, we can represent these outcomes by the Cartesian product $C \times d_1$, which is a horizontal segment at d_1. The probability that the vector $\left[\begin{smallmatrix}\tilde{c}\\ \tilde{d}\end{smallmatrix}\right]$ belongs to $C \times d_1$ equals $\mathrm{P}(D_1 \cap C^{-1})$, represented by the area of $D_1 \cap C^{-1}$ in the Venn diagram.

The following theorem shows how to compute the marginal pdf of a continuous random variable from its conditional pdf given a discrete random variable. We sum the conditional pdfs weighted by the pmf of the discrete random variable.

Theorem 6.2. *Let $F_{\tilde{c}\,|\,\tilde{d}}$ and $f_{\tilde{c}\,|\,\tilde{d}}$ be the conditional cdf and pdf of a continuous*

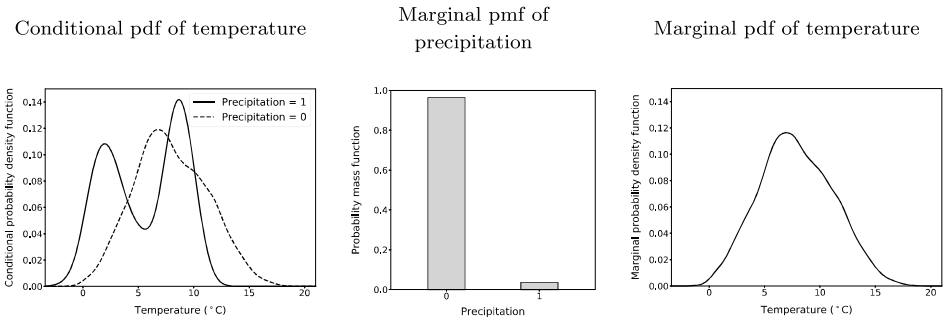

Figure 6.2 Joint distribution of precipitation and temperature in Mauna Loa. The left graph shows the conditional pdf of the temperature at the Mauna Loa weather station (Hawaii) in 2015 depending on whether it rains or snows (precipitation = 1) or not (precipitation =0), obtained via kernel density estimation. The center graph shows the marginal pmf of precipitation; there is precipitation only about 3.6% of the time. The right graph shows the marginal pdf of temperature, which is equal to the sum of the conditional pdfs weighted by the entries of the pmf, as established in Theorem 6.2. As a result, the marginal pdf is very close to the conditional pdf given no precipitation, because the corresponding value of the pmf is much higher.

random variable $\tilde{c}$ given a discrete random variable $\tilde{d}$, with range $R_{\tilde{d}}$. Then,

$$F_{\tilde{c}}(c) = \sum_{d \in R_{\tilde{d}}} p_{\tilde{d}}(d)\, F_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d)\,, \tag{6.6}$$

$$f_{\tilde{c}}(c) = \sum_{d \in R_{\tilde{d}}} p_{\tilde{d}}(d)\, f_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d)\,. \tag{6.7}$$

Proof The events $\{\tilde{d} = d\}$ are a partition of the whole probability space (one of them must happen and they are all disjoint), so by the law of total probability

$$F_{\tilde{c}}(c) = \mathrm{P}\left(\tilde{c} \leq c\right) \tag{6.8}$$

$$= \sum_{d \in R_{\tilde{d}}} \mathrm{P}\left(\tilde{d} = d\right) \mathrm{P}\left(\tilde{c} \leq c \,|\, \tilde{d} = d\right) \tag{6.9}$$

$$= \sum_{d \in R_{\tilde{d}}} p_{\tilde{d}}(d)\, F_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d)\,. \tag{6.10}$$

Now, (6.7) follows by differentiating. ■

Combining a discrete marginal pmf with a continuous conditional distribution allows us to define parametric *mixture models* where the data is drawn from a parametric distribution whose parameters depend on the value of a discrete quantity. If a Gaussian is used as the continuous distribution, this yields a Gaussian mixture model. In Sections 6.5 and 6.6, we describe applications of Gaussian mixture models to the tasks of classification and clustering, respectively.

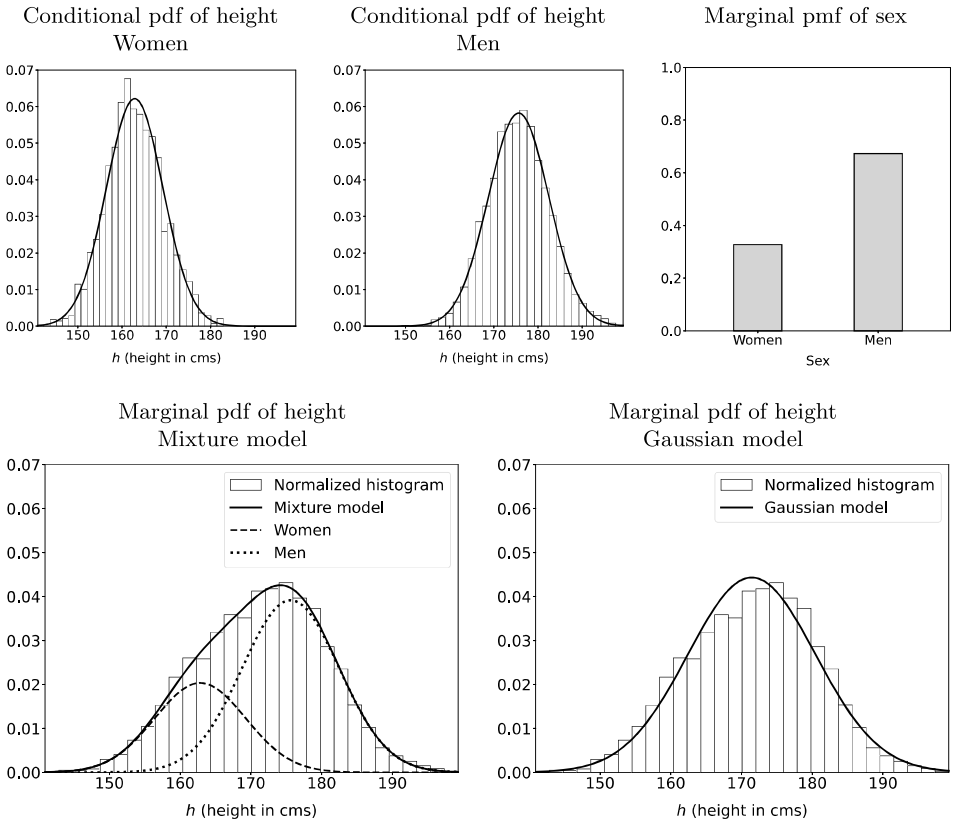

Figure 6.3 Gaussian mixture model for height. The top row shows a Gaussian parametric model for the height of 1,986 women (left) and 4,082 men (center) in the US army. The marginal pmf of sex is shown on the right. The left graph of the bottom row shows the marginal pdf obtained by applying Theorem 6.2 as explained in Example 6.3. This is a Gaussian mixture model that provides a substantially better approximation to the data than the Gaussian model shown on the right, obtained by fitting a Gaussian model to all individuals regardless of sex.

Example 6.3 (Gaussian mixture model for height)**.** In Figure 3.23 we show the result of fitting a Gaussian parametric model to height data from 4,082 men in the United States army, extracted from Dataset 5. If we instead fit the Gaussian model to the whole population, which also includes of 1,986 women then the fit is not as good, as shown in the lower right plot of Figure 6.3. However, we can model the height of the women as a Gaussian distribution, as long as we fit it separately (see top left graph in Figure 6.3). Motivated by this, we represent height as a continuous random variable $\tilde{h}$, sex as a discrete random variable $\tilde{s}$, and we model the conditional pdf of height given sex as Gaussian. We fit the conditional Gaussian pdfs via maximum-likelihood estimation (see Section 3.7). For women,

the mean is $\mu_{\text{women}} := 163$ cm and the standard deviation is $\sigma_{\text{women}} := 6.4$ cm. For men, the mean is $\mu_{\text{men}} := 176$ cm and the standard deviation is $\sigma_{\text{men}} := 6.9$ cm. The random variable $\tilde{s}$ is Bernoulli: $\tilde{s} = 0$ and $\tilde{s} = 1$ indicate that the individual is a woman or a man respectively. The maximum-likelihood estimate of the Bernoulli parameter is $\alpha_{\text{men}} := 0.67$, because 67% of the people are men (see Example 2.26). By (6.7) the pdf of $\tilde{h}$ is

$$
\begin{aligned}
&f_{\tilde{h}}(h) \\
&= \sum_{s=0}^{1} p_{\tilde{s}}(s)\, f_{\tilde{h}\,|\,\tilde{s}}(h\,|\,s) \qquad (6.11)\\
&= \frac{\alpha_{\text{women}}}{\sqrt{2\pi}\sigma_{\text{women}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\text{women}}}{\sigma_{\text{women}}}\right)^2\right) + \frac{\alpha_{\text{men}}}{\sqrt{2\pi}\sigma_{\text{men}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\text{men}}}{\sigma_{\text{men}}}\right)^2\right)\\
&= \frac{0.33}{6.4\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{h-163}{6.4}\right)^2\right) + \frac{0.67}{6.9\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{h-176}{6.9}\right)^2\right), \qquad (6.12)
\end{aligned}
$$

where $\alpha_{\text{men}} := 1 - \alpha_{\text{women}}$. Figure 6.3 shows the conditional pdfs of $\tilde{h}$ and the marginal pmf of $\tilde{s}$ on the top row. The bottom row shows that the mixture model (left) produces a better fit than if we use a Gaussian model for all the data (right).

6.3 Conditional Distribution Of Discrete Variables

Recall that the probability that a continuous random variable $\tilde{c}$ is equal to a specific value c equals zero (see Section 3.1). Consequently, it is not immediately obvious how to define the conditional pmf of a discrete random variable $\tilde{d}$ given $\tilde{c} = c$. This is the same issue we encounter in Section 5.6 when defining the conditional pdf. We follow the same approach and define the conditional pmf as a limit. The conditional pmf of $\tilde{d}$ at d given $\tilde{c} = c$ is the conditional probability of $\tilde{d} = d$ given that $\tilde{c}$ is in an interval that includes c, in the limit when the length of the interval shrinks to zero.

Definition 6.4 (Conditional pmf of a discrete random variable given a continuous random variable)**.** *Let $\tilde{c}$ and $\tilde{d}$ be a continuous and a discrete random variable defined on the same probability space. The conditional pmf of $\tilde{d}$ given $\tilde{c}$ is*

$$
p_{\tilde{d}\,|\,\tilde{c}}(d\,|\,c) := \lim_{\epsilon \to 0} \mathrm{P}\left(\tilde{d} = d \,|\, c - \epsilon < \tilde{c} \leq c\right). \qquad (6.13)
$$

Analogously to Theorem 6.2, if we have access to the conditional pmf $p_{\tilde{d}\,|\,\tilde{c}}$ of a discrete random variable $\tilde{d}$ given a continuous random variable $\tilde{c}$, we can obtain the marginal pmf of $\tilde{d}$ by integrating $p_{\tilde{d}\,|\,\tilde{c}}$ weighted by the marginal pdf of $\tilde{c}$.

Theorem 6.5. *Let $p_{\tilde{d}\,|\,\tilde{c}}$ be the conditional pmf of a discrete random variable $\tilde{d}$ given a continuous random variable $\tilde{c}$. Then,*

$$
p_{\tilde{d}}(d) = \int_{c=-\infty}^{\infty} f_{\tilde{c}}(c)\, p_{\tilde{d}\,|\,\tilde{c}}(d\,|\,c)\, \mathrm{d}c. \qquad (6.14)
$$

Proof We do not give a formal proof but rather an intuitive argument that can be made rigorous. If we take a grid of values for c which are on a grid $\{..., c_{-1}, c_0, c_1, ...\}$ with step size ϵ, then

$$p_{\tilde{d}}(d) = \sum_{i=-\infty}^{\infty} \mathrm{P}\left(\tilde{d} = d, c_i - \epsilon < \tilde{c} \leq c_i\right) \tag{6.15}$$

by the law of total probability. Taking the limit as $\epsilon \to 0$, the sum becomes an integral, so that

$$p_{\tilde{d}}(d) = \int_{c=-\infty}^{\infty} \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(\tilde{d} = d, c - \epsilon < \tilde{c} \leq c\right)}{\epsilon} \mathrm{d}c \tag{6.16}$$

$$= \int_{c=-\infty}^{\infty} \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(c - \epsilon < \tilde{c} \leq c\right)}{\epsilon} \cdot \mathrm{P}\left(\tilde{d} = d \,|\, c - \epsilon < \tilde{c} \leq c\right) \mathrm{d}c \tag{6.17}$$

$$= \int_{c=-\infty}^{\infty} f_{\tilde{c}}(c)\, p_{\tilde{d}\,|\,\tilde{c}}(d \,|\, c)\, \mathrm{d}c. \tag{6.18}$$

■

The following theorem provides an analog to the chain rule for jointly distributed discrete and continuous random variables.

Theorem 6.6 (Chain rule for discrete and continuous random variables)**.** *Let $\tilde{c}$ be a continuous random variable with conditional pdf $f_{\tilde{c}\,|\,\tilde{d}}$ and $\tilde{d}$ a discrete random variable with conditional pmf $p_{\tilde{d}\,|\,\tilde{c}}$. Then,*

$$p_{\tilde{d}}(d)\, f_{\tilde{c}\,|\,\tilde{d}}(c \,|\, d) = f_{\tilde{c}}(c)\, p_{\tilde{d}\,|\,\tilde{c}}(d \,|\, c)\,. \tag{6.19}$$

Proof By the definitions of the conditional pmf and of the pdf,

$$p_{\tilde{d}}(d)\, f_{\tilde{c}\,|\,\tilde{d}}(c \,|\, d) = \lim_{\epsilon \to 0} \mathrm{P}(\tilde{d} = d) \frac{\mathrm{P}\left(c - \epsilon < \tilde{c} \leq c \,|\, \tilde{d} = d\right)}{\epsilon} \tag{6.20}$$

$$= \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(\tilde{d} = d, c - \epsilon < \tilde{c} \leq c\right)}{\epsilon} \tag{6.21}$$

$$= \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(c - \epsilon < \tilde{c} \leq c\right)}{\epsilon} \mathrm{P}\left(\tilde{d} = d \,|\, c - \epsilon < \tilde{c} \leq c\right) \tag{6.22}$$

$$= f_{\tilde{c}}(c)\, p_{\tilde{d}\,|\,\tilde{c}}(d \,|\, c)\,. \tag{6.23}$$

■

Interpreted as a function of c and d,

$$p_{\tilde{d}}(d)\, f_{\tilde{c}\,|\,\tilde{d}}(c \,|\, d) = f_{\tilde{c}}(c)\, p_{\tilde{d}\,|\,\tilde{c}}(d \,|\, c) \tag{6.24}$$

captures the local joint probabilistic behavior of the two random variables, and can therefore be interpreted as a surrogate for their joint pmf and pdf, which are not well defined. Figure 6.4 shows this function for the Gaussian mixture model in Example 6.3.

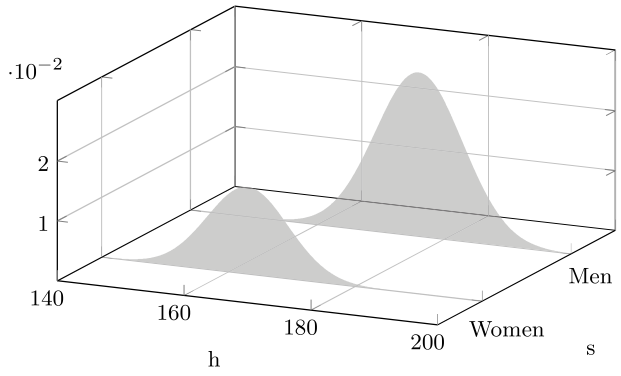

Figure 6.4 Describing the joint distribution of discrete and continuous random variables. The product of the marginal pmf and the conditional pdf (or equivalently, of the marginal pdf and the conditional pmf) plays an analogous role to the joint pmf of discrete random variables, or the joint pdf of continuous random variables. The graph shows the function $p_{\tilde{s}}(s)\, f_{\tilde{h} \mid \tilde{s}}(h \mid s) = f_{\tilde{h}}(h)\, p_{\tilde{s} \mid \tilde{h}}(s \mid h)$ for Example 6.3.

An important application of Theorem 6.6 is to determine the conditional distribution of a discrete random variable $\tilde{d}$ given a continuous random variable $\tilde{c}$ from the marginal distributions of $\tilde{d}$ and $\tilde{c}$, and the continuous distributions of $\tilde{c}$ given $\tilde{d}$:

$$p_{\tilde{d} \mid \tilde{c}}(d \mid c) = \frac{p_{\tilde{d}}(d)\, f_{\tilde{c} \mid \tilde{d}}(c \mid d)}{\sum_{d \in R_{\tilde{d}}} p_{\tilde{d}}(d)\, f_{\tilde{c} \mid \tilde{d}}(c \mid d)}, \tag{6.25}$$

where $R_{\tilde{d}}$ is the range of $\tilde{d}$ and the expression in the denominator is a consequence of Theorem 6.2. This is useful to perform classification, as we explain in Section 6.5. The left plot of Figure 6.5 shows the conditional pmf of precipitation given temperature at the Mauna Loa weather station, computed by applying (6.25) to the conditional pdfs and marginal pmf estimates shown in Figure 6.2. In the case of the Gaussian mixture model from Example 6.3, the conditional probability of an individual being a woman given that their height is

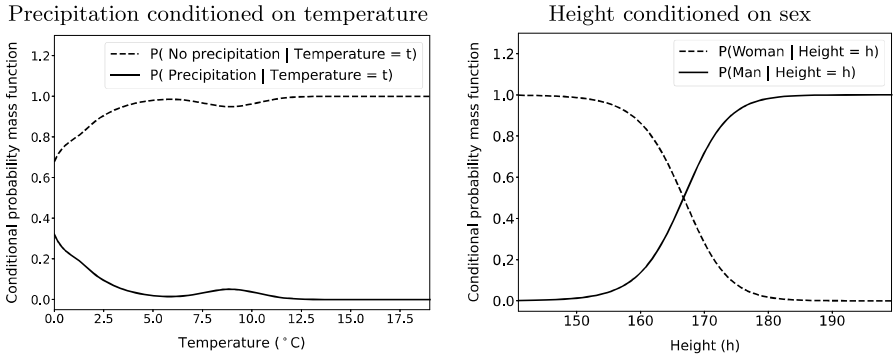

Figure 6.5 Conditional distribution of discrete variables given continuous variables. The left graph shows the conditional pmf of precipitation at the Mauna Loa weather station (Hawaii) in 2015 given the temperature, computed via (6.25) using the pmf and conditional pdfs from Figure 6.2. The probability of precipitation is higher at low temperatures. The right graph shows the conditional pmf of sex given height in the US army, derived in (6.30). The probability of woman is very high for small heights, and then undergoes a sharp transition, becoming very low for large heights.

h equals

$$p_{\tilde{s}\,|\,\tilde{h}}\,(0\,|\,h) \tag{6.26}$$

$$= \frac{p_{\tilde{s}}\,(0)\,f_{\tilde{h}\,|\,\tilde{s}}\,(h\,|\,0)}{p_{\tilde{s}}\,(0)\,f_{\tilde{h}\,|\,\tilde{s}}\,(h\,|\,0) + p_{\tilde{s}}\,(1)\,f_{\tilde{h}\,|\,\tilde{s}}\,(h\,|\,1)} \tag{6.27}$$

$$= \frac{\frac{\alpha_{\text{women}}}{\sqrt{2\pi}\sigma_{\text{women}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\text{women}}}{\sigma_{\text{women}}}\right)^2\right)}{\frac{\alpha_{\text{women}}}{\sqrt{2\pi}\sigma_{\text{women}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\text{women}}}{\sigma_{\text{women}}}\right)^2\right) + \frac{\alpha_{\text{men}}}{\sqrt{2\pi}\sigma_{\text{men}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\text{men}}}{\sigma_{\text{men}}}\right)^2\right)} \tag{6.28}$$

$$= \frac{1}{1 + \frac{\alpha_{\text{men}}}{\alpha_{\text{women}}}\frac{\sigma_{\text{women}}}{\sigma_{\text{men}}} \exp\left(\frac{1}{2}\left(\frac{h-\mu_{\text{women}}}{\sigma_{\text{women}}}\right)^2 - \frac{1}{2}\left(\frac{h-\mu_{\text{men}}}{\sigma_{\text{men}}}\right)^2\right)} \tag{6.29}$$

$$= \frac{1}{1 + 0.7\exp\left(0.0017h^2 - 0.28h\right)}. \tag{6.30}$$

The right plot of Figure 6.5 shows the conditional pmf. At small heights, the probability of woman is almost one. As we increase the height, the probability of man increases, with a quick transition between 160 and 170 cm. The shape of the function that maps height to the corresponding conditional pmf looks like a logistic function. The following theorem shows that if the variances of the conditional distributions are equal, then the function is indeed a logistic curve.

Theorem 6.7 (Logistic function)**.** *Let $\tilde{d}$ be a Bernoulli random variable with parameter α and let $\tilde{c}$ be a continuous random variable defined on the same prob-*

ability space. If the conditional distributions of $\tilde{c}$ given $\tilde{d} = 0$ and $\tilde{d} = 1$ are both Gaussian with mean parameters μ_0 and μ_1, respectively, and the same standard-deviation parameter σ, then the conditional pmf of $\tilde{d}$ given $\tilde{c} = c$ equals

$$p_{\tilde{d}\,|\,\tilde{c}}(1\,|\,c) = \frac{1}{1+\eta\exp(-\beta c)}, \tag{6.31}$$

which is a logistic function of c where

$$\eta := \frac{1-\alpha}{\alpha}\exp\left(\frac{1}{2\sigma^2}\left(\mu_1^2-\mu_0^2\right)\right), \tag{6.32}$$

$$\beta := \frac{\mu_1-\mu_0}{\sigma^2}. \tag{6.33}$$

Proof By (6.25) and the assumptions,

$$p_{\tilde{d}\,|\,\tilde{c}}\,(1\,|\,c) = \frac{p_{\tilde{d}}\,(1)\,f_{\tilde{c}\,|\,\tilde{d}}\,(c\,|\,1)}{p_{\tilde{d}}\,(0)\,f_{\tilde{c}\,|\,\tilde{d}}\,(c\,|\,0)+p_{\tilde{d}}\,(1)\,f_{\tilde{c}\,|\,\tilde{d}}\,(c\,|\,1)} \tag{6.34}$$

$$= \frac{\frac{\alpha}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{c-\mu_1}{\sigma}\right)^2\right)}{\frac{\alpha}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{c-\mu_1}{\sigma}\right)^2\right)+\frac{1-\alpha}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{c-\mu_0}{\sigma}\right)^2\right)} \tag{6.35}$$

$$= \frac{1}{1+\frac{1-\alpha}{\alpha}\exp\left(\frac{1}{2}\left(\frac{c-\mu_1}{\sigma}\right)^2-\frac{1}{2}\left(\frac{c-\mu_0}{\sigma}\right)^2\right)} \tag{6.36}$$

$$= \frac{1}{1+\frac{1-\alpha}{\alpha}\exp\left(\frac{\mu_0-\mu_1}{\sigma^2}c+\frac{1}{2\sigma^2}\left(\mu_1^2-\mu_0^2\right)\right)}. \tag{6.37}$$

■

To visualize the logistic function derived in Theorem 6.7, let us assume that the two values of $\tilde{d}$ are equally likely ($\alpha := 0.5$), and that the conditional distributions are centered at -1 ($\mu_0 := -1$) and 1 ($\mu_1 := 1$) and have unit variance ($\sigma := 1$). The conditional probability of $\tilde{d} = 1$ given $\tilde{c} = c$ then corresponds to the logistic function depicted in Figure 12.21 which is almost 0 when c is negative, and then transitions to 1 as c increases, crossing 0.5 exactly at the origin.

6.4 Independence

For a continuous random variable $\tilde{a}$ and a discrete random variable $\tilde{d}$, we define independence in a similar way as for discrete and continuous random variables: the conditional distributions are equal to the marginal distribution.

Definition 6.8 (Independence)**.** *A pair of discrete and continuous random variables $\tilde{d}$ and $\tilde{c}$ defined on the same probability space are independent if and only if*

$$p_{\tilde{d}\,|\,\tilde{c}}(d\,|\,c) = p_{\tilde{d}}(d), \tag{6.38}$$

$$f_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d) = f_{\tilde{c}}(c) \tag{6.39}$$

for all possible values of c and d.

We define conditional independence for discrete and continuous random variables analogously.

Definition 6.9 (Conditionally independent random variables)**.** *A pair of discrete and continuous random variables $\tilde{d}$ and $\tilde{c}$ defined on the same probability space are conditionally independent given a random variable $\tilde{a}$ if and only if*

$$p_{\tilde{d}\,|\,\tilde{c},\tilde{a}}(d\,|\,c,a) = p_{\tilde{d}\,|\,\tilde{a}}(d\,|\,a), \tag{6.40}$$

$$f_{\tilde{c}\,|\,\tilde{d},\tilde{a}}(c\,|\,d,a) = f_{\tilde{c}\,|\,\tilde{a}}(c\,|\,a), \tag{6.41}$$

for all values of a, c and d.

Figure 6.6 shows the conditional and marginal pdfs and pmfs of height and handedness (whether their left or right hand is dominant) for members of the US army, extracted from Dataset 5. The conditional and marginal distributions are very similar, which suggests that height and handedness are independent.

6.5 Classification Via Gaussian Discriminant Analysis

In Section 4.8 we consider the task of classifying data into predefined classes based on discrete features. Here, we present an approach to perform classification from continuous features. As a motivating application, we consider the automatic diagnosis of Alzheimer's disease, a neurodegenerative disease that causes dementia. We assume that we have available n training examples (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n). Each example consists of a d-dimensional vector x_i of features and its corresponding label, denoted by $y_i \in \{1, 2, \ldots, c\}$, where c is the number of classes. In Alzheimer's diagnosis, there are two classes: healthy individuals ($y_i = 0$) and patients with Alzheimer's ($y_i = 1$). We consider a simple set of features consisting of the volumes of the hippocampus and entorhinal cortex, two regions of the brain that shrink due to Alzheimer's. The volumes are measured using magnetic resonance imaging and normalized by the intracranial volume of the head (to take into account that some people have larger heads than others). Our goal is to diagnose whether someone has Alzheimer's from these features.

In order to tackle the classification task, we interpret each example as a sample from the joint distribution of a d-dimensional random vector $\tilde{x}$ representing the features, and a random variable $\tilde{y}$ representing the corresponding class label. We can then apply *maximum a posteriori* estimation, as in Section 4.8, to classify each new example based on the associated feature vector x. The MAP estimator is obtained by maximizing the conditional pmf of $\tilde{y}$ given $\tilde{x} = x$, just as in the discrete case (see Definition 4.29).

Definition 6.10 (MAP estimator)**.** *Given a discrete random variable $\tilde{y}$ with range Y and a continuous random vector $\tilde{x}$, the maximum a posteriori (MAP) estimator of $\tilde{y}$ given $\tilde{x} = x$ is*

$$\text{MAP}(x) := \arg\max_{y \in Y} p_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x). \tag{6.42}$$

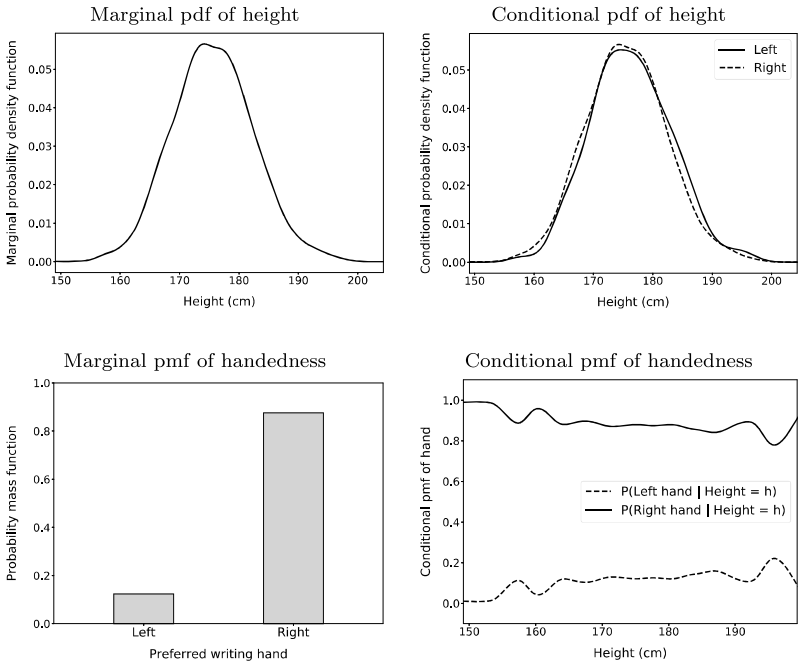

Figure 6.6 Joint distribution of height and handedness. The top right graph shows the conditional pdf of height given a dominant left or right hand (for writing), obtained via kernel density estimation (KDE). The top left graph shows the marginal pdf of the height, also estimated via KDE. The conditional pdfs and the marginal pdf are very similar, suggesting that height and handedness are independent. The bottom right plot shows the conditional pmf of handedness given height, obtained by applying Theorem 6.6. The bottom left plot shows the marginal pmf of handedness; about 11.5% of the individuals are left handed. The conditional pmf is approximately constant and equal to the marginal, except at small and large heights, where the number of data is small.

Just as in classification from discrete features (see Theorem 4.30), the MAP estimator is optimal in the sense that it minimizes the probability of error, and hence maximizes the accuracy of the estimator.

Theorem 6.11 (MAP estimation is optimal)**.** *Given a discrete random variable $\tilde{y}$ and a d-dimensional continuous random vector $\tilde{x}$, the maximum a posteriori (MAP) estimate of $\tilde{y}$ given $\tilde{x} = x$, minimizes the probability of error. Equivalently, the probability that the MAP estimator is correct is greater than or equal to the probability that any estimator $h : \mathbb{R}^d \to \mathbb{R}$ is correct:*

$$\mathrm{P}\left(\mathrm{MAP}(\tilde{x}) = \tilde{y}\right) \geq \mathrm{P}\left(h(\tilde{x}) = \tilde{y}\right). \tag{6.43}$$

Proof To prove the result, we apply the same argument in the proof of The-

orem 4.30 to the case where $\tilde{x}$ is continuous. In order to do this, we need to apply the law of total probability to the union of events with zero probability $\tilde{x} = x$, for all possible values of x. By the same limit argument as in the proof of Theorem 6.5,

$$\begin{aligned}
\mathrm{P}\left(h(\tilde{x}) = \tilde{y}\right) &= \int_{x[1]=-\infty}^{\infty} \cdots \int_{x[d]=-\infty}^{\infty} f_{\tilde{x}}(x) P\left(\tilde{y} = h\left(x\right) \mid \tilde{x} = x\right) \, \mathrm{d}x[1] \ldots \, \mathrm{d}x[d] \\
&= \int_{x[1]=-\infty}^{\infty} \cdots \int_{x[d]=-\infty}^{\infty} f_{\tilde{x}}(x) p_{\tilde{y} \mid \tilde{x}}\left(h\left(x\right) \mid x\right) \, \mathrm{d}x[1] \ldots \, \mathrm{d}x[d] \qquad (6.44) \\
&\leq \int_{x[1]=-\infty}^{\infty} \cdots \int_{x[d]=-\infty}^{\infty} f_{\tilde{x}}(x) p_{\tilde{y} \mid \tilde{x}}\left(\mathrm{MAP}\left(x\right) \mid x\right) \, \mathrm{d}x[1] \ldots \, \mathrm{d}x[d] \qquad (6.45) \\
&= \int_{x[1]=-\infty}^{\infty} \cdots \int_{x[d]=-\infty}^{\infty} f_{\tilde{x}}(x) P\left(\tilde{y} = \mathrm{MAP}\left(x\right) \mid \tilde{x} = x\right) \, \mathrm{d}x[1] \ldots \, \mathrm{d}x[d] \\
&= \mathrm{P}\left(\mathrm{MAP}(\tilde{x}) = \tilde{y}\right), \qquad (6.46)
\end{aligned}$$

where the inequality holds because, by the definition of the MAP estimate, $p_{\tilde{y} \mid \tilde{x}}\left(\mathrm{MAP}\left(x\right) \mid x\right) \geq p_{\tilde{y} \mid \tilde{x}}\left(h\left(x\right) \mid x\right)$ for any value of x. ■

Unfortunately, the MAP estimator is intractable to compute unless there are very few features due to the curse of dimensionality. The number of possible values of the feature vector $\tilde{x}$ scales exponentially with its dimension d, as explained in Section 4.7, so it is often impossible to estimate all the possible conditional pmfs of $\tilde{y}$ given $\tilde{x}$ from the training data. This is the same phenomenon described in Section 4.8, where we circumvent the curse of dimensionality by making *naive* conditional independence assumptions that render MAP estimation tractable. Here we take a different route: we assume that the distribution of the features $\tilde{x}$ follows a *parametric model*, with parameters that depend on the class $\tilde{y}$. Specifically, we model $\tilde{x}$ as a Gaussian mixture model where the discrete random variable is the class $\tilde{y}$. We fit this parametric model from data and use the resulting distributions to estimate the conditional pmfs $p_{\tilde{y} \mid \tilde{x}}$. This classification method is known as Gaussian discriminant analysis.

Definition 6.12 (Gaussian discriminant analysis)**.** *Let (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) be a dataset of n examples, where x_i denotes a d-dimensional continuous vector and $y_i \in \{1, 2, \ldots, c\}$ its corresponding class (c is the number of classes). To perform Gaussian discriminant analysis, we first fit a Gaussian mixture model to the training data. The pmf of $\tilde{y}$ is estimated using the corresponding empirical pmf (see Definition 2.11). The resulting estimate for $p_{\tilde{y}}(k)$ is*

$$\alpha_k := \frac{n_k}{n}, \quad k \in \{1, 2, \ldots, c\}, \qquad (6.47)$$

where n_k is the number of examples labeled as belonging to class k. Let us denote these examples by $(x_1^{[k]}, y_1^{[k]})$, $(x_2^{[k]}, y_2^{[k]})$, $\ldots$, $(x_{n_k}^{[k]}, y_{n_k}^{[k]})$. We model the conditional pdf of $\tilde{x}$ given $\tilde{y} = k$ as a Gaussian random vector, obtaining the corresponding mean and covariance-matrix parameters via maximum-likelihood estimation (see

Section 5.10.3),

$$\mu_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i^{[k]}, \tag{6.48}$$

$$\Sigma_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_i^{[k]} - \mu_k)(x_i^{[k]} - \mu_k)^T. \tag{6.49}$$

To classify a new example, we compute the MAP estimate of $\tilde{y}$ given $\tilde{x} = x$ by applying (6.25)

$$\text{MAP}\,(x) := \arg \max_{y \in \{1,2,\ldots,c\}} p_{\tilde{y} \mid \tilde{x}}(y \mid x) \tag{6.50}$$

$$= \arg \max_{y \in \{1,2,\ldots,c\}} \frac{p_{\tilde{y}}(y) f_{\tilde{x} \mid \tilde{y}}(x \mid y)}{\sum_{k \in \{1,2,\ldots,c\}} p_{\tilde{y}}(k) p_{\tilde{x} \mid \tilde{y}}(x \mid k)} \tag{6.51}$$

$$= \arg \max_{y \in \{1,2,\ldots,c\}} \frac{\frac{\alpha_y}{\sqrt{(2\pi)^d |\Sigma_y|}} \exp\left(-\frac{1}{2} (x - \mu_y)^T \Sigma_y^{-1} (x - \mu_y)\right)}{\sum_{k \in \{1,2,\ldots,c\}} \frac{\alpha_k}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left(-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)}. \tag{6.52}$$

In order to fit a Gaussian mixture model, we need to estimate a d-dimensional mean and a $d \times d$ covariance matrix for each class, as well as the marginal pmf of each class. The number of parameters scales *quadratically* with the number of features, instead of exponentially, which makes it tractable to fit the model, as long as the number of features is not extremely large.

Figure 6.7 shows the result of applying Gaussian discriminant analysis to perform diagnosis of Alzheimer's disease using real-world data extracted from Dataset 12. The plots show the contour lines of the two Gaussian conditional distributions obtained by fitting the training data (top row), and of the estimated conditional probability of Alzheimer's given the data (middle row). If we diagnose Alzheimer's when the conditional probability is above 0.5, the resulting accuracy (the fraction of correct diagnoses) is 75.8%. For comparison, a naive classifier that never diagnoses Alzheimer's attains an accuracy of 72.9%, because the fraction of healthy subjects in the dataset is 72.9%. Note, however, that in healthcare applications, accuracy is often not the preferred evaluation metric. Section 12.9 provides an in-depth discussion of how to evaluate classification models.

A critical consideration in real-world applications is how models *generalize* to new data points. To evaluate generalization, we use a test dataset consisting of a different population of individuals. These test examples are shown on the bottom row of Figure 6.7. Comparing our training and test sets reveals a crucial challenge that often arises in practice: there is a systematic difference between the distribution of the features in the training and test sets. This is often referred to as *domain shift*. In this case, our method is robust to the shift: the model yields an accuracy of 81.5% on the test data. In comparison, the fraction of healthy

subjects in the test population is 78.4%, so we again achieve an improvement of around 3% over the naive baseline.

Gaussian discriminant analysis is often also called *quadratic discriminant analysis* in the literature. To see why, let us consider the values of the feature vector x for which the conditional probability of two classes a and b are equal,

$$\frac{p_{\tilde{y}\,|\,\tilde{x}}(a\,|\,x)}{p_{\tilde{y}\,|\,\tilde{x}}(b\,|\,x)} = 1. \tag{6.53}$$

The resulting hypersurface separates the points for which a is more likely than b according to the model, so we think of it as a *decision boundary*. By (6.52), the boundary is given by the equation

$$\frac{\alpha_a\sqrt{|\Sigma_b|}}{\alpha_b\sqrt{|\Sigma_a|}} \exp\left(\frac{1}{2}\left(x-\mu_b\right)^T\Sigma_b^{-1}\left(x-\mu_b\right) - \frac{1}{2}\left(x-\mu_a\right)^T\Sigma_a^{-1}\left(x-\mu_a\right)\right) = 1.$$

Taking logarithms, we obtain a quadratic hypersurface of dimension $d-1$ (where d is the dimension of the features),

$$\frac{1}{2}\left(x-\mu_b\right)^T\Sigma_b^{-1}\left(x-\mu_b\right) - \frac{1}{2}\left(x-\mu_a\right)^T\Sigma_a^{-1}\left(x-\mu_a\right) + \log\left(\frac{\alpha_a\sqrt{|\Sigma_b|}}{\alpha_b\sqrt{|\Sigma_a|}}\right) = 0. \tag{6.54}$$

The middle left plot of Figure 6.7 shows this quadratic decision boundary for the Alzheimer's example, where the boundary is a one-dimensional curve because $d = 2$.

Gaussian discriminant analysis can be adapted to yield a linear decision boundary by imposing the constraint that the covariance-matrix parameters $\Sigma_1, \ldots, \Sigma_c$ in Definition 6.12 are all equal to the same matrix Σ. The matrix is estimated via maximum-likelihood estimation using all the data,

$$\Sigma := \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{y_i})(x_i - \mu_{y_i})^T. \tag{6.55}$$

Notice that each data point is centered using the mean parameter corresponding to its respective class. Setting both covariance matrices equal to Σ in (6.54) yields

$$\beta^T x + \xi = 0, \tag{6.56}$$

where

$$\beta := \Sigma^{-1}\left(\mu_a - \mu_b\right), \tag{6.57}$$

$$\xi := \frac{1}{2}\mu_b^T\Sigma^{-1}\mu_b - \frac{1}{2}\mu_a^T\Sigma^{-1}\mu_a + \log\frac{\alpha_a}{\alpha_b}, \tag{6.58}$$

which is indeed linear in x, so the decision boundary is a hyperplane. This version of Gaussian discriminant analysis is called *linear discriminant analysis*.

The right column of Figure 6.7 shows the result of applying linear discriminant analysis to our Alzheimer's example. The top plot shows the contour lines of the two Gaussian conditional distributions, which are identical shifted copies of one

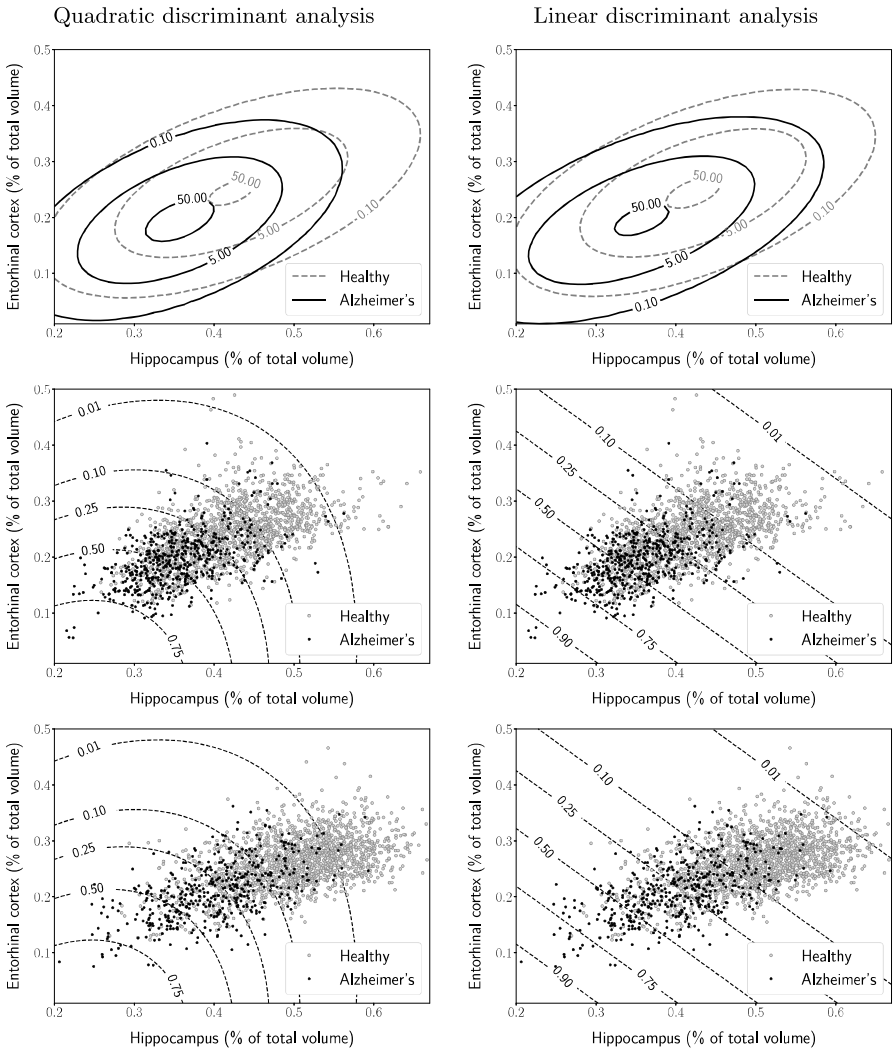

Figure 6.7 Diagnosis of Alzheimer's disease via Gaussian discriminant analysis. The top row shows the contour lines of Gaussian distributions obtained by fitting normalized volumes of the hippocampus and the entorhinal cortex extracted from 1,926 magnetic-resonance scans in the Alzheimer's Disease Neuroimaging Initiative dataset (**?**). The data are separated into two groups: healthy individuals (gray) and patients with Alzheimer's disease (black). The middle row shows the probability of Alzheimer's computed via (6.25), superposed on the training data. The bottom row shows the same probability superposed on a test dataset corresponding to 2,046 magnetic-resonance scans from the National Alzheimer's Coordinating Center dataset (**?**), used as a test set. The left column corresponds to quadratic-discriminant-analysis model where each Gaussian distribution has a different covariance matrix. The right column corresponds to a linear-discriminant-analysis model where the Gaussians have the same covariance matrix.

another. The contour lines of the conditional probability of Alzheimer's given the data are shown in the middle and bottom plots, superposed on the training and test data, respectively. The decision boundaries are indeed linear. The accuracy is 75.9% for the training data and 81.5% for the test data, essentially the same as that of quadratic discriminant analysis. For these features, it may be difficult to outperform a simple linear classifier. In order to improve performance, we need to incorporate additional features into our model.

6.6 Clustering Via Gaussian Mixture Models

6.6.1 Latent Variables

Consider the height data in Example 6.3. In order to design a mixture model appropriate for these data, we exploit the fact that they can be separated into two groups according to sex. In this section we consider the problem of fitting such a model *without knowing what group each data point belongs to*. To this end, we define a *latent* or *hidden* discrete variable associated to each data point, and infer its value from the data. This is called *unsupervised* learning, as opposed to the *supervised* setting of Section 6.5, where the discrete class associated to each data point is known. An important application of mixture models is clustering, which is a fundamental task in exploratory data analysis. The goal is to separate the data into groups or *clusters* of similar points. This can be achieved by fitting a mixture model and then using the learned latent variable to determine the clusters.

We focus our discussion on Gaussian mixture models, which is the most popular mixture model used for clustering, but the same ideas can be adapted to other parametric distributions. Our objective is to model a dataset of d-dimensional real-valued vectors $x_1, x_2, \ldots, x_n$ as samples from a continuous random variable $\tilde{x}$. The distribution of $\tilde{x}$ depends on a latent discrete random variable $\tilde{k}$ with range $\{1, 2, \ldots, m\}$, which is not observed. If $\tilde{k} = k$, then $\tilde{x}$ belongs to the kth cluster and its conditional distribution is Gaussian with parameters that depend on k. We assume that the number of clusters m is fixed beforehand.

6.6.2 Fitting Gaussian Mixture Models With Latent Variables

In order to fit the Gaussian mixture model described in Section 6.6.1 to data, we need to estimate the pmf of $\tilde{k}$ and the parameters of all the conditional Gaussian distributions of $\tilde{x}$ given $\tilde{k}$. As explained in Section 3.7, a reasonable approach to fit the parameters of a parametric model is to maximize its likelihood, which equals its probability density at the observed data, usually under an i.i.d. assumption. To ease notation, let us denote the parameters of the Gaussian mixture model by

$$\theta := \left\{ \{\alpha_k\}_{k=1}^m, \{\mu_k\}_{k=1}^m, \{\Sigma_k\}_{k=1}^m \right\}. \tag{6.59}$$

The parameter $\alpha_k \in [0, 1]$ represents the probability that $\tilde{k} = k$, so it must satisfy $\sum_{k=1}^m \alpha_k = 1$. The parameters $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d \times d}$ are the mean

and covariance-matrix parameters of $\tilde{x}$ given $\tilde{k} = k$, respectively, so Σ_k must be symmetric and positive definite (see Definition 5.21). The conditional likelihood of each data point i given $\tilde{k} = k$ is

$$\mathcal{L}_{i,k}(\mu_k, \Sigma_k) := f_{\tilde{x} \,|\, \tilde{k}}(x_i \,|\, k) \tag{6.60}$$

$$= \frac{1}{\sqrt{(2\pi)^d \, |\Sigma_k|}} \exp\left(-\frac{1}{2} \left(x_i - \mu_k\right)^T \Sigma_k^{-1} \left(x_i - \mu_k\right) \right). \tag{6.61}$$

If the data are i.i.d., by Theorem 6.2, the likelihood of the whole dataset equals

$$\mathcal{L}(\theta) := \prod_{i=1}^{n} f_{\tilde{x}}(x_i) \tag{6.62}$$

$$= \prod_{i=1}^{n} \sum_{k=1}^{m} p_{\tilde{k}}(k) f_{\tilde{x} \,|\, \tilde{k}}(x_i \,|\, k) \tag{6.63}$$

$$= \prod_{i=1}^{n} \sum_{k=1}^{m} \alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k), \tag{6.64}$$

and the corresponding log-likelihood is

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{m} \alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k). \tag{6.65}$$

Up to now, whenever we have encountered a log-likelihood function, it has been relatively straightforward to derive the parameter values that maximize it (see Sections 2.4, 3.7 or 5.10.3). This is not the case for the log-likelihood in (6.65). In particular, it is not concave and has several local maxima. As an example, when $m = 2$, if there is a maximum at $\alpha_1 := \alpha_a$, $\mu_1 := \mu_a$, $\Sigma_1 := \Sigma_a$, $\alpha_2 = 1 - \alpha_a$, $\mu_2 := \mu_b$, $\Sigma_2 := \Sigma_b$ for some $\alpha_a \in [0, 1]$ and some μ_a, Σ_a, μ_b and Σ_b with appropriate dimensions, then there must also be a maximum at $\alpha_1 := 1 - \alpha_a$, $\mu_1 := \mu_b$, $\Sigma_1 := \Sigma_b$, $\alpha_2 = \alpha_a$, $\mu_2 := \mu_a$, $\Sigma_2 := \Sigma_a$, because the log-likelihood attains exactly the same value.

Our strategy to fit Gaussian mixture models is to maximize the log-likelihood iteratively, until we reach a local maximum. This can be achieved in different ways, but by far the most popular is the *expectation-maximization* algorithm. The algorithm jointly fits the parameters and an auxiliary quantity $\gamma_{i,k}$, which we call *membership probabilities*. The kth membership probability associated to the ith data point x_i is the conditional probability that x_i belongs to the kth cluster given $\tilde{x} = x_i$. By (6.25),

$$\gamma_{i,k} := p_{\tilde{k} \,|\, \tilde{x}}\left(k \,|\, x_i\right) \tag{6.66}$$

$$= \frac{p_{\tilde{k}}\left(k\right) f_{\tilde{x_i} \,|\, \tilde{k}}\left(x_i \,|\, k\right)}{\sum_{l=1}^{m} p_{\tilde{k}}\left(l\right) f_{\tilde{x_i} \,|\, \tilde{k}}\left(x_i \,|\, l\right)} \tag{6.67}$$

$$= \frac{\alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k)}{\sum_{l=1}^{m} \alpha_l \mathcal{L}_{i,l}(\mu_l, \Sigma_l)}. \tag{6.68}$$

This quantity is critical for clustering. Once we fit the mixture model, we can assign point i to a cluster by choosing the most likely cluster according to the model,

$$\hat{k}_i := \arg\max_k p_{\tilde{k} \mid \tilde{x}}\left(k \mid x_i\right) \tag{6.69}$$

$$= \arg\max_k \gamma_{i,k}. \tag{6.70}$$

The sum of $\gamma_{i,k}$ over all data points can be interpreted as the *effective cardinality* (number of points) of the kth cluster,

$$n_k := \sum_{i=1}^{n} \gamma_{i,k}. \tag{6.71}$$

The expectation-maximization algorithm alternates between updating the membership probabilities in the *expectation* step, and the parameters α_k, μ_k and Σ_k for each cluster k in the *maximization* step. It can be shown that each update of the expectation-maximization algorithm increases the log-likelihood of the model (we refer the reader to Section 9.4 in (**?**) for a proof).

Section 6.6.3 provides a formal justification of the updates in the expectation-maximization algorithm, but first let us motivate them intuitively. Imagine that we know what cluster each data point corresponds to. Then we should set α_k equal to the fraction of data points in cluster k, and μ_k and Σ_k equal to the sample mean and sample covariance matrix of the data points in that cluster (since that is the corresponding maximum-likelihood estimate by Theorem 5.26). Of course, we don't know the cluster assignments, but if we have an estimate of the membership probabilities, we can use them as a *soft assignment*: We estimate α_k as the ratio between the effective cardinality and the total number of points, and μ_k and Σ_k as the sample mean and sample covariance matrix *weighted by the membership probabilities*, so that the contribution of each point depends on the probability that it belongs to the cluster. This yields the updates (6.74), (6.75) and (6.76) in Definition 6.13.

Definition 6.13 (Expectation maximization for Gaussian mixture models)**.** *Let $x_1, x_2, \ldots, x_n$ be n d-dimensional real-valued vectors. To fit a Gaussian mixture models with m clusters to these data, we first initialize the parameters $\{\alpha_k\}_{k=1}^m$, $\{\mu_k\}_{k=1}^m$ and $\{\Sigma_k\}_{k=1}^m$, ensuring that $\alpha_k \in [0,1]$, $\sum_{k=1}^m \alpha_k = 1$, $\mu_k \in \mathbb{R}^d$ and every $\Sigma_k \in \mathbb{R}^{d \times d}$ is symmetric and positive definite. This can be done by separating the points into m groups (either randomly or using some other clustering approach such as k-means), and applying (6.74), (6.75) and (6.76). We then repeat the following steps until the log-likelihood converges:*

1 Update the membership probabilities $\gamma_{i,k}$ for $1 \leq i \leq n$, $1 \leq k \leq m$, and also

the corresponding effective cluster cardinality n_k

$$\gamma_{i,k} := \frac{\alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k)}{\sum_{l=1}^{m} \alpha_l \mathcal{L}_{i,l}(\mu_l, \Sigma_l)}, \tag{6.72}$$

$$n_k := \sum_{i=1}^{n} \gamma_{i,k}. \tag{6.73}$$

2 Update the parameters of each Gaussian conditional distribution

$$\alpha_k := \frac{n_k}{n}, \tag{6.74}$$

$$\mu_k := \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} x_i, \tag{6.75}$$

$$\Sigma_k := \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} \left(x_i - \mu_k\right) \left(x_i - \mu_k\right)^T. \tag{6.76}$$

After convergence, we assign a cluster $\hat{k}_i \in \{1, \ldots, m\}$ *to each data point by setting*

$$\hat{k}_i := \arg\max_k \gamma_{i,k}. \tag{6.77}$$

Example 6.14 (Clustering according to height)**.** In this example, we fit a Gaussian mixture model with two clusters to the height data in Example 6.3, without using the fact that we know the sex of each individual. Figure 6.8 shows the convergence of the log-likelihood over the iterations of the expectation-maximization algorithm, as well as the convergence of the parameters in the mixture model. Figure 6.9 shows the corresponding mixture distribution and the conditional probabilities of each cluster given the data for some of the iterations. The model parameters eventually converge to provide a good fit to the data.

The two clusters identified by the model make a lot of sense. Even though the model has no access to sex information, the mean (175 cm) and standard deviation (7.4 cm) of the first cluster are close to the mean (176 cm) and standard deviation (6.9 cm) of the men. Likewise, the mean (161 cm) and standard deviation (5.8 cm) of the second cluster are close to the mean (163 cm) and standard deviation (6.4 cm) of the women. The Gaussian mixture model is able to automatically identify the two subgroups in the population.

Example 6.15 (Clustering basketball players)**.** In this example, we apply a Gaussian mixture model to analyze basketball statistics. The top left plot in Figure 6.10 shows a scatterplot of the assists and rebounds per game of NBA players between 1996 and 2019, extracted from Dataset 13. We apply a Gaussian mixture model with three clusters to the data, obtaining the conditional Gaussian distributions shown on the top right. Cluster 1 corresponds to players for which the number of assists and rebounds per game are very small. The two other clusters are more interesting.

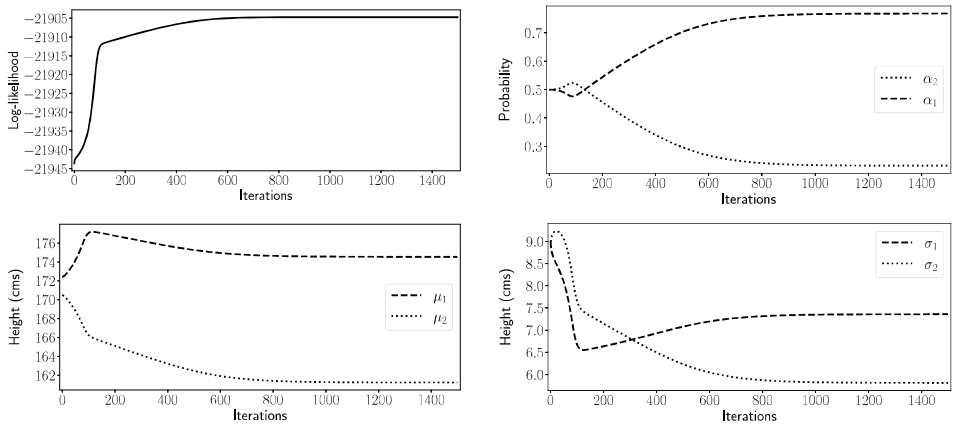

Figure 6.8 Convergence of the expectation-maximization algorithm to fit a Gaussian mixture model. The top left graph shows the evolution of the log-likelihood (6.65) when we apply expectation maximization to fit a Gaussian mixture model with two clusters to the height data in Example 6.3. The log-likelihood increases monotonically and eventually converges. The remaining plots show the convergence of the parameters α_1 and α_2 (top right), μ_1 and μ_2 (bottom left), and σ_1 and σ_2 (bottom right).

Cluster 2 consists of players that catch a lot of rebounds, but do not assist very much. In contrast, cluster 3 consists of players who assist a lot, but do not catch many rebounds. The model is able to automatically group players with similar responsibilities. This is confirmed by the height distributions of the players assigned to each cluster, shown on the bottom of Figure 6.10 (note that we are *not* using this information to produce the clusters). The heights of the players in cluster 2 are typical of centers and forwards, whose role is indeed typically to catch rebounds rather than pass the ball (with some exceptions like Nikola Jokic). The heights of the players in cluster 3 are typical of guards, who usually give more assists, and catch less rebounds.

6.6.3 Derivation Of The Expectation-Maximization Updates

Our strategy to fit the Gaussian mixture model is to try to maximize the log-likelihood $\log \mathcal{L}(\theta)$ defined in (6.65) by finding parameter values for which its gradient is zero. Unfortunately, setting the gradient to zero does not yield equations that we can solve in closed form (in contrast to the models in Sections 2.4, 3.7 or 5.10.3). In order to make some progress, recall that maximizing the Gaussian conditional log-likelihood $\mathcal{L}_{i,k}(\mu_k, \Sigma_k)$, $1 \leq k \leq m$, $1 \leq i \leq n$, with respect to μ_k and Σ_k is easy (see Section 5.10.3). Inspired by this, we express the log-likelihood

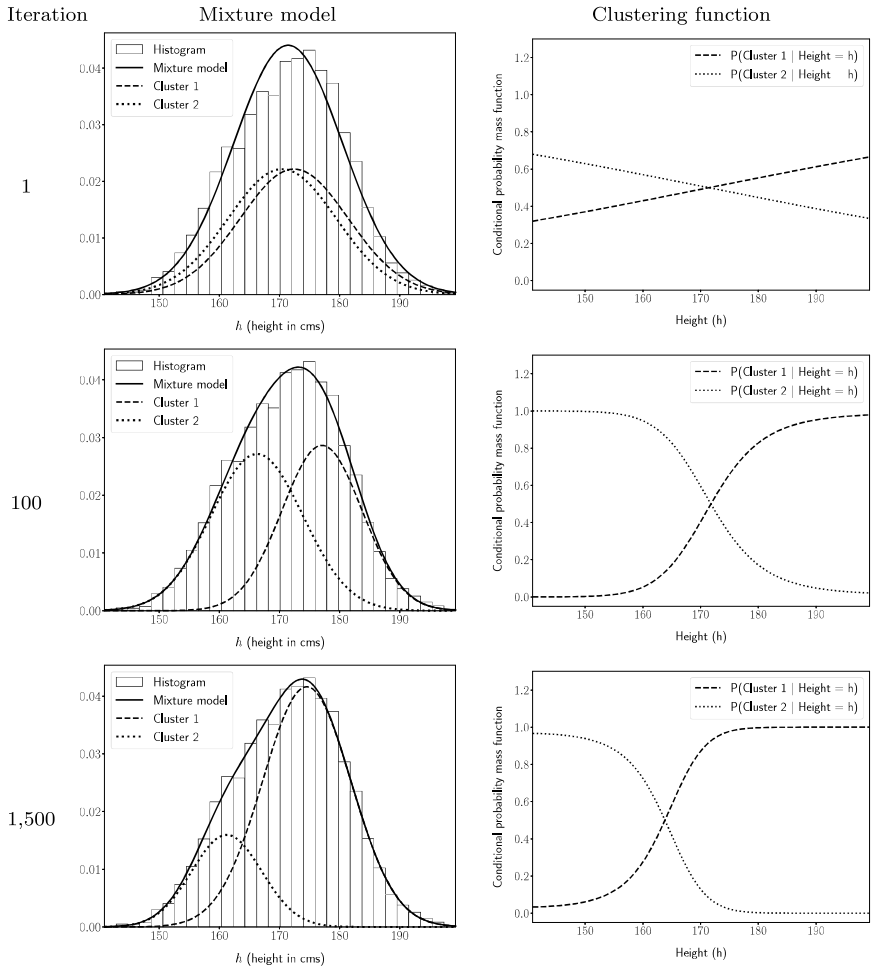

Figure 6.9 Gaussian mixture model for height. The plots show the results of fitting a Gaussian mixture model with two clusters to the height data in Example 6.3. The left column shows the probability density of the model (solid line) compared to the data (represented as a histogram), as well as the contributions of each cluster to the density at different iterations of the expectation-maximization algorithm. The fit clearly improves as the iterations proceed. The right column shows the conditional probabilities of each cluster given the data at the same iterations. Eventually, the conditional pdf separates the data in a very similar way to the conditional pmf of sex given height in the right plot of Figure 6.5, even though the mixture model does not have access to sex information.

as a function of the conditional log-likehood:

$$\nabla_{\mu_k,\Sigma_k} \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \frac{\alpha_k \nabla_{\mu_k,\Sigma_k} \mathcal{L}_{i,k}(\mu_k, \Sigma_k)}{\sum_{l=1}^{m} \alpha_l \mathcal{L}_{i,l}(\mu_l, \Sigma_l)} \tag{6.78}$$

$$= \sum_{i=1}^{n} \frac{\alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k)}{\sum_{l=1}^{m} \alpha_l \mathcal{L}_{i,l}(\mu_l, \Sigma_l)} \nabla_{\mu_k,\Sigma_k} \log \mathcal{L}_{i,k}(\mu_k, \Sigma_k) \tag{6.79}$$

$$= \sum_{i=1}^{n} \gamma_{i,k} \nabla_{\mu_k,\Sigma_k} \log \mathcal{L}_{i,k}(\mu_k, \Sigma_k), \tag{6.80}$$

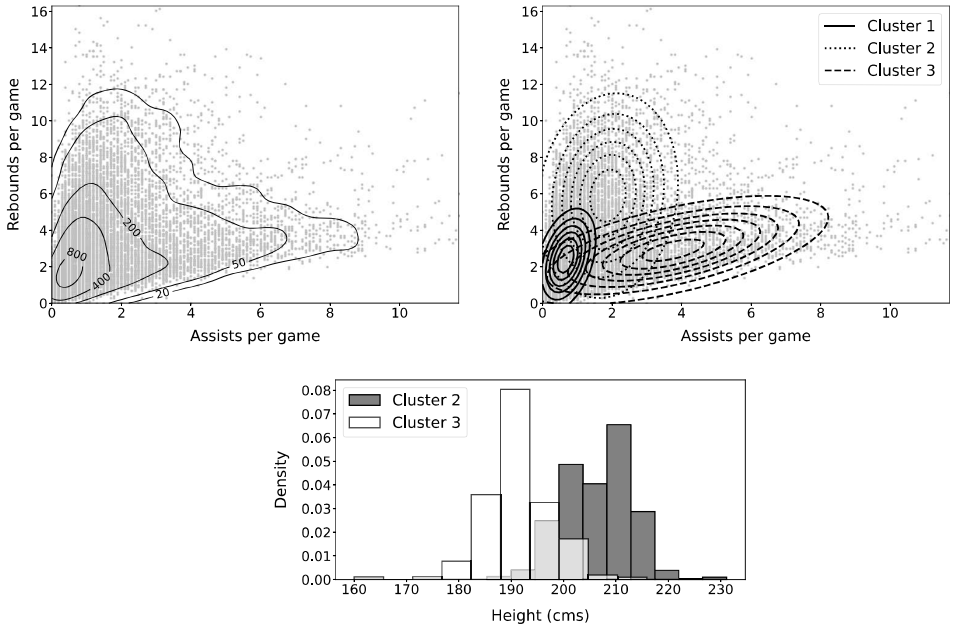

Figure 6.10 Gaussian mixture model for basketball players. The top left plot shows a scatterplot of the assists and rebounds per game of NBA players between 1996 and 2019, as well as the corresponding probability density estimated via multidimensional kernel density estimation (see Section 5.4). The top right plot shows the Gaussian densities associated to each of the three clusters identified by a Gaussian mixture model fit to the data using expectation maximization. The bottom plot shows histograms of the heights of players assigned to clusters 2 and 3. The players in cluster 2 are taller than those in cluster 3, which shows that the clustering algorithm automatically separates the players into guards, who are typically shorter, and centers and forwards, who are taller.

where we have used the fact that

$$\nabla_{\mu_k,\Sigma_k} \log \mathcal{L}_{i,k}(\mu_k, \Sigma_k) = \mathcal{L}_{i,k}(\mu_k, \Sigma_k)^{-1} \nabla_{\mu_k,\Sigma_k} \mathcal{L}_{i,k}(\mu_k, \Sigma_k). \tag{6.81}$$

If all the $\gamma_{i,k}$ were constant with respect to the model parameters, then it would be easy to maximize this expression with respect to μ_k and Σ_k. Unfortunately, $\gamma_{i,k}$ is a function of μ_k, Σ_k and the rest of model parameters. The key idea of the expectation-maximization algorithm is to ignore this, and derive simple updates for the model parameters assuming that $\gamma_{i,k}$ is constant for all $1 \leq i \leq n$, $1 \leq k \leq m$.

Under the assumption that $\gamma_{i,k}$ is fixed, the gradient of the log-likelihood with respect to μ_k and Σ_k for any $1 \leq k \leq m$ can be obtained following the same

derivations as in the proof of Theorem 5.26. We have

$$\sum_{i=1}^{n} \gamma_{i,k} \nabla_{\mu_k} \log \mathcal{L}_{i,k}(\mu_k, \Sigma_k) = \sum_{i=1}^{n} \gamma_{i,k} \Sigma_k^{-1} (x_i - \mu_k) \tag{6.82}$$

$$= \Sigma_k^{-1} \sum_{i=1}^{n} \gamma_{i,k} (x_i - \mu_k), \tag{6.83}$$

$$\sum_{i=1}^{n} \gamma_{i,k} \nabla_{\Sigma_k^{-1}} \log \mathcal{L}_{i,k}(\mu_k, \Sigma_k) = \sum_{i=1}^{n} \frac{\gamma_{i,k}}{2} \left(\Sigma_k - (x_i - \mu_k)(x_i - \mu_k)^T \right). \tag{6.84}$$

Setting these expressions to zero yields

$$\mu_k^* = \frac{\sum_{i=1}^{n} \gamma_{i,k} x_i}{\sum_{i=1}^{n} \gamma_{i,k}} \tag{6.85}$$

$$= \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} x_i, \tag{6.86}$$

$$\Sigma_k^* = \frac{\sum_{i=1}^{n} \gamma_{i,k} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} \gamma_{i,k}} \tag{6.87}$$

$$= \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} (x_i - \mu_k)(x_i - \mu_k)^T. \tag{6.88}$$

In order to determine the update for α_k, $1 \leq k \leq m$, we need to enforce the constraint that $\sum_{k=1}^{m} \alpha_k = 1$. This is achieved by setting $\alpha_m := 1 - \sum_{k=1}^{m-1} \alpha_k$, so that the log-likelihood equals

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{m-1} \alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k) + \left(1 - \sum_{k=1}^{m-1} \alpha_k \right) \mathcal{L}_{i,m}(\mu_m, \Sigma_m) \right). \tag{6.89}$$

The partial derivative of the log-likelihood with respect to α_k is

$$\frac{\partial \log \mathcal{L}(\theta)}{\partial \alpha_k} = \sum_{i=1}^{n} \frac{\mathcal{L}_{i,k}(\mu_k, \Sigma_k) - \mathcal{L}_{i,m}(\mu_m, \Sigma_m)}{\sum_{l=1}^{m} \alpha_l \mathcal{L}_{i,l}(\mu_l, \Sigma_l)}. \tag{6.90}$$

Multiplying the partial derivative by α_k yields

$$\alpha_k \frac{\partial \log \mathcal{L}(\theta)}{\partial \alpha_k} = \sum_{i=1}^{n} \frac{\alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k) - \alpha_k \mathcal{L}_{i,m}(\mu_m, \Sigma_m)}{\sum_{l=1}^{m} \alpha_l \mathcal{L}_{i,l}(\mu_l, \Sigma_l)} \tag{6.91}$$

$$= \sum_{i=1}^{n} \gamma_{i,k} - \frac{\alpha_k \mathcal{L}_{i,m}(\mu_m, \Sigma_m)}{\sum_{l=1}^{m} \alpha_l \mathcal{L}_{i,l}(\mu_l, \Sigma_l)} \tag{6.92}$$

$$= n_k - \frac{\alpha_k}{\alpha_m} \sum_{i=1}^{n} \frac{\alpha_m \mathcal{L}_{i,m}(\mu_m, \Sigma_m)}{\sum_{l=1}^{m} \alpha_l \mathcal{L}_{i,l}(\mu_l, \Sigma_l)} \tag{6.93}$$

$$= n_k - \frac{\alpha_k}{\alpha_m} n_m. \tag{6.94}$$

We now assume that $\gamma_{i,k}$ is constant for $1 \leq i \leq n$, $1 \leq k \leq m$, and therefore so

is n_k (but remember that this is not necessarily true!), and find the values of α_k for which the partial derivatives vanish. We sum (6.94) over k and set it equal to zero:

$$\sum_{k=1}^{m} n_k - \sum_{k=1}^{m} \frac{\alpha_k}{\alpha_m} n_m = n - \frac{n_m}{\alpha_m} = 0, \tag{6.95}$$

where we have used the fact that $\sum_{k=1}^{m} n_k = n$ and $\sum_{k=1}^{m} \alpha_k = 1$. We conclude that α_m should equal $\frac{n_m}{n}$. Plugging this into (6.94) yields $n_k - \alpha_k n$, which equals zero if α_k equals

$$\alpha_k^* := \frac{n_k}{n}, \quad 1 \leq k \leq m. \tag{6.96}$$

We check that indeed $\sum_{k=1}^{m} \alpha_k^* = 1$.

To summarize, we have derived values for the mixture-model parameters for which the gradient of the log-likelihood vanishes if $\gamma_{i,k}$ is constant for all $1 \leq i \leq n$, $1 \leq k \leq m$. Of course, there is a catch. As we already mentioned, $\gamma_{i,k}$ depends on the parameters! In order for the gradient of the log-likelihood to actually vanish, we also need the following equation to hold

$$\gamma_{i,k} = \frac{\alpha_k^* \mathcal{L}_{i,k}(\mu_k^*, \Sigma_k^*)}{\sum_{k=1}^{m} \alpha_l^* \mathcal{L}_{i,l}(\mu_l^*, \Sigma_l^*)}. \tag{6.97}$$

In order to resolve this chicken-and-egg-problem, we take a pragmatic approach. We alternate between (1) updating μ_k, Σ_k and α_k using (6.86), (6.88) and (6.96), and (2) updating $\gamma_{i,k}$ via (6.97). Our hope is that the alternating scheme will eventually converge to a value for the parameters so that all the equations hold. Step 1 is known as the *maximization* step because it maximizes the log-likelihood with respect to the parameters. Step 2 is known as the *expectation* step, because one can interpret the log-likelihood function as the expected log-likelihood with respect to the membership probabilities. Putting everything together, we obtain the algorithm in Definition 6.13. Figures 6.8 and 6.9 show the expectation-maximization updates converging to a local maximum of the log-likelihood. Even after convergence, we are not guaranteed to have reached a global maximum, because the function is nonconvex. However, we must remember that our ultimate goal is obtaining a *useful model*, and Gaussian mixture models often uncover useful clusterings in real data, as illustrated in Examples 6.14 and 6.15.

6.7 Bayesian Parametric Modeling

6.7.1 The Bayesian Framework

In traditional parametric modeling, described in Sections 2.3 and 3.6, we model data using predefined distributions that depend on a small number of parameters. These parameters are deterministic quantities, chosen so that the parametric model approximates the data as accurately as possible (e.g. via maximum likelihood). In contrast, Bayesian parametric modeling represents the parameters as

random variables. This allows us to quantify our uncertainty about these parameters in a principled way. Building a Bayesian model parametric requires making two important decisions. We need to choose:

- The marginal distribution of the parameters $\tilde{\theta}$, in the form of their marginal pmf $p_{\tilde{\theta}}$ or marginal pdf $f_{\tilde{\theta}}$. We call this the *prior* distribution of the parameters, because it captures our uncertainty *before* seeing the data. The prior can be used to incorporate assumptions that reflect any additional information we have about the parameters.
- The conditional distribution of the data $\tilde{x}$ given the parameters, in the form of the conditional pmf $p_{\tilde{x}\,|\,\tilde{\theta}}$ or the conditional pdf $f_{\tilde{x}\,|\,\tilde{\theta}}$. The conditional distribution is called the *likelihood* of the data. For a fixed value of the parameters, this is the same likelihood we define in Sections 2.4 and 3.7; it represents the probability or the probability density of the data. The crucial difference is that in traditional parametric modeling, the likelihood has no probabilistic meaning, it is just interpreted as a function of the parameters. In Bayesian modeling, it is a valid conditional pmf or conditional pdf.

The main goal in Bayesian parametric modeling is to compute the conditional distribution of the parameters $\tilde{\theta}$ given the observed data x, in the form of the conditional pmf $p_{\tilde{\theta}\,|\,\tilde{x}}$ or the conditional pdf $f_{\tilde{\theta}\,|\,\tilde{x}}$ evaluated at $\tilde{x} = x$. This is called the *posterior* distribution of the parameters, because it represents our uncertainty *after* seeing the data.

Example 6.16 (Single coin flip)**.** We consider the problem of modeling a single coin flip, using a parametric Bernoulli model with a parameter that represents the probability of heads. In traditional parametric modeling, the parameter is a deterministic variable θ. The likelihood equals

$$\mathcal{L}\left(\theta\right) = \begin{cases} \theta & \text{if flip is heads,} \\ 1-\theta & \text{if flip is tails.} \end{cases} \tag{6.98}$$

The maximum likelihood estimate obtained by maximizing the likelihood is

$$\theta_{\text{ML}} = \begin{cases} 1 & \text{if flip is heads,} \\ 0 & \text{if flip is tails.} \end{cases} \tag{6.99}$$

Under the traditional framework, there is no uncertainty associated to the parameter. In contrast, under a Bayesian framework, we model the parameter as a random variable $\tilde{\theta}$. We consider two possible marginal pdfs for $\tilde{\theta}$: a uniform pdf in $[0, 1]$, and a triangular pdf of the form

$$f_{\tilde{\theta}}\left(\theta\right) = 2\theta \quad \text{for } \theta \in [0, 1] \ . \tag{6.100}$$

The triangular pdf encodes the assumption that the coin is more likely to land on heads.

Conditioned on $\tilde{\theta} = \theta$, the result of the coin flip $\tilde{x}$ is modeled as a Bernoulli

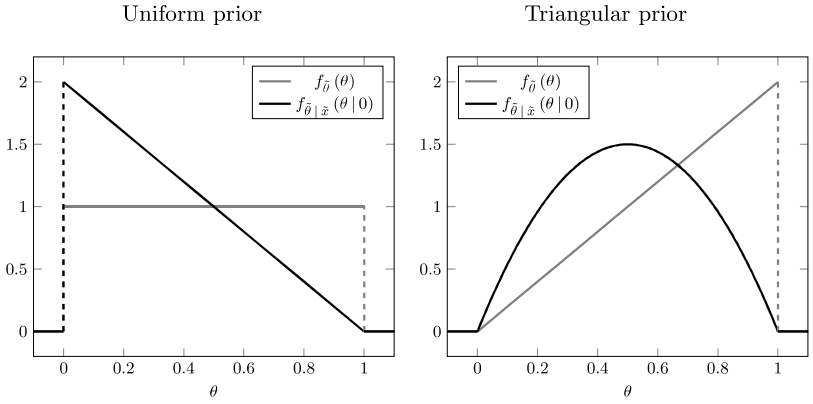

Figure 6.11 Prior and posterior distributions for the coin flip in Example 6.16. The prior on the left does not make any assumptions about the fairness of the coin. The prior on the right encodes the assumption that the coin may be more prone to land on heads. The corresponding posterior pdfs update the uncertainty about the Bernoulli parameter that governs the coin flip after we observe that it lands on tails.

random variable with parameter θ ($\tilde{x} = 1$ is heads and $\tilde{x} = 0$ is tails). The likelihood is the conditional pmf of $\tilde{x}$ given θ:

$$p_{\tilde{x}\,|\,\tilde{\theta}}(x\,|\,\theta) = \begin{cases} \theta & \text{if } x = 1, \\ 1-\theta & \text{if } x = 0. \end{cases} \tag{6.101}$$

It is exactly the same as the likelihood (6.98) of the traditional parametric model, but it has *a very different interpretation*: it is a conditional probability mass function, and not just a function of a deterministic parameter.

Imagine that the coin lands on tails, so $\tilde{x} = 0$. The posterior pdf of the parameter given this data point depends on the prior that we choose. For the uniform prior it equals

$$f_{\tilde{\theta}\,|\,\tilde{x}}\,(\theta\,|\,0) = \frac{f_{\tilde{\theta}}\,(\theta)\,p_{\tilde{x}\,|\,\tilde{\theta}}\,(0\,|\,\theta)}{p_{\tilde{x}}\,(0)} \tag{6.102}$$

$$= \frac{1-\theta}{\int_{u=-\infty}^{\infty} f_{\tilde{\theta}}\,(u)\,p_{\tilde{x}\,|\,\tilde{\theta}}\,(0\,|\,u)\;\mathrm{d}u} \tag{6.103}$$

$$= \frac{1-\theta}{\int_{u=0}^{1} 1-u\;\mathrm{d}u} \tag{6.104}$$

$$= 2\,(1-\theta)\,. \tag{6.105}$$

The posterior pdf is concentrated near zero, indicating that maybe the probability of tails is higher than the probability of heads. For the triangular prior, the

posterior pdf equals

$$f_{\tilde{\theta}\,|\,\tilde{x}}\,(\theta\,|\,0) = \frac{f_{\tilde{\theta}}\,(\theta)\,p_{\tilde{x}\,|\,\tilde{\theta}}\,(0\,|\,\theta)}{p_{\tilde{x}}\,(0)} \tag{6.106}$$

$$= \frac{2\theta\,(1-\theta)}{\int_{u=-\infty}^{\infty} f_{\tilde{\theta}}\,(u)\,p_{\tilde{x}\,|\,\tilde{\theta}}\,(0\,|\,u)\;\mathrm{d}u} \tag{6.107}$$

$$= \frac{2\theta\,(1-\theta)}{\int_{u=0}^{1} 2u\,(1-u)\;\mathrm{d}u} \tag{6.108}$$

$$= 6\theta\,(1-\theta)\,. \tag{6.109}$$

The posterior pdf is centered, suggesting that the coin may not be biased towards heads after all. Figure 6.11 shows the prior and posterior pdfs for the two priors. Due to the small number of data (just one!), both posterior pdfs are very spread out.

Example 6.16 just considers a single data point, which results in pretty inconclusive posterior distributions. If more data are available, we encounter another modeling choice: characterizing the joint distribution of the random variables representing the data. A common assumption is that the data are conditionally independent given the parameters. This is analogous to the i.i.d. assumption in Sections 2.4 and 3.7, where the data are assumed to be independent and to share a common parameter. Let $\tilde{\theta}$ represent the parameters and let $\tilde{x}$ be an n-dimensional random vector representing the data. If we assume conditional independence, the likelihood of the data equals

$$p_{\tilde{x}\,|\,\tilde{\theta}}(x\,|\,\theta) = \prod_{i=1}^{n} p_{\tilde{x}[i]\,|\,\tilde{\theta}}(x[i]\,|\,\theta), \tag{6.110}$$

if the data are discrete, or

$$f_{\tilde{x}\,|\,\tilde{\theta}}(x\,|\,\theta) = \prod_{i=1}^{n} f_{\tilde{x}[i]\,|\,\tilde{\theta}}(x[i]\,|\,\theta), \tag{6.111}$$

if the data are continuous.

6.7.2 The Beta Distribution And Conjugate Priors

In discrete parametric models, model parameters often encode a probability. This is the case for the Bernoulli, geometric and binomial models presented in Section 2.3. The beta distribution is a useful parametric model for the prior of such parameters. It provides a family of continuous probability densities on the unit interval. The two parameters a and b determine the shape of the distribution, as illustrated in Figure 6.12. The priors and posterior distributions in Example 6.16 are all beta distributions.

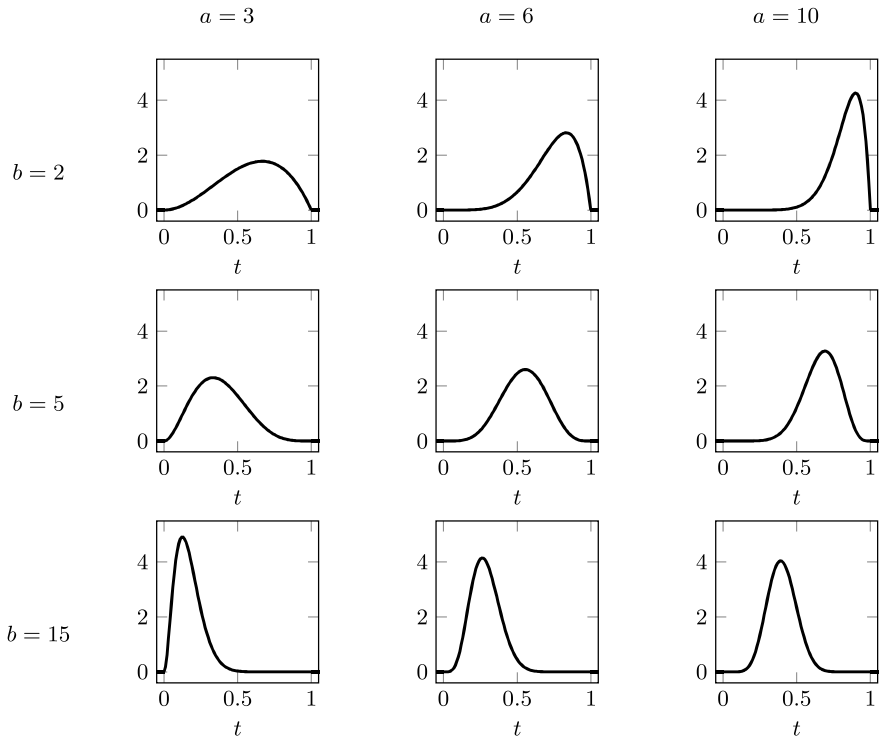

Figure 6.12 Beta distribution. For the examples above, the probability density of the beta distribution has a single maximum whose location is determined by the value of the a and b parameters. When a is larger than b, the maximum is located towards the right of the unit interval. When we increase b, it shifts to the left. Increasing both a and b results in a pdf that is more concentrated around the maximum.

Definition 6.17 (Beta distribution)**.** *The pdf of a beta random variable $\tilde{t}$ with parameters a and b is*

$$f_{\tilde{t}}(t) := \frac{t^{a-1}(1-t)^{b-1}}{\beta(a,b)} \qquad \text{if} \quad 0 \leq t \leq 1, \tag{6.112}$$

and zero otherwise, where

$$\beta(a,b) := \int_u u^{a-1}(1-u)^{b-1}\,\mathrm{d}u \tag{6.113}$$

is a special function called the beta function or Euler integral of the first kind, which must be computed numerically.

Let us consider the problem of estimating the parameter of a Bernoulli distribution from n data points that are conditionally independent given the Bernoulli

parameter. Under the conditional independence assumption, the data have a binomial distribution. As established in the following theorem, if we choose a beta prior for the Bernoulli parameter, then the posterior is also a beta distribution! In Bayesian modeling, we say that the beta distribution is a *conjugate prior* to the binomial likelihood.

Theorem 6.18. *Let the continuous random variable $\tilde{\theta}$ represent a parameter restricted to the unit interval. If the prior distribution of $\tilde{\theta}$ is a beta distribution with parameters a and b and the likelihood of the data $\tilde{x}$ given $\tilde{\theta} = \theta$ is binomial with parameters n and θ, then the posterior distribution of $\tilde{\theta}$ given $\tilde{x} = x$ is a beta distribution with parameters $x + a$ and $n - x + b$.*

Proof By Definition 6.17,

$$f_{\tilde{\theta} \mid \tilde{x}}(\theta \mid x) = \frac{f_{\tilde{\theta}}(\theta)\, p_{\tilde{x} \mid \tilde{\theta}}(x \mid \theta)}{p_{\tilde{x}}(x)} \tag{6.114}$$

$$= \frac{f_{\tilde{\theta}}(\theta)\, p_{\tilde{x} \mid \tilde{\theta}}(x \mid \theta)}{\int_{u=0}^{1} f_{\tilde{\theta}}(u)\, p_{\tilde{x} \mid \tilde{\theta}}(x \mid u)\, \mathrm{d}u} \tag{6.115}$$

$$= \frac{\theta^{a-1}(1-\theta)^{b-1}\binom{n}{x}\theta^{x}(1-\theta)^{n-x}}{\int_{u=0}^{1} u^{a-1}(1-u)^{b-1}\binom{n}{x}u^{x}(1-u)^{n-x}\,\mathrm{d}u} \tag{6.116}$$

$$= \frac{\theta^{x+a-1}(1-\theta)^{n-x+b-1}}{\int_{u=0}^{1} u^{x+a-1}(1-u)^{n-x+b-1}\,\mathrm{d}u}, \tag{6.117}$$

which is the pdf of a beta random variable with parameters $x+a$ and $n-x+b$. ■

In the following example, we apply Bayesian modeling based on the beta distribution to analyze poll data from the 2020 election.

Example 6.19 (Poll in Pennsylvania)**.** In a poll in Pennsylvania for the 2020 US election, 281 people intend to vote for Trump and 300 for Biden. The data are from a real poll (**?**), where we ignore other candidates and undecided voters for simplicity. We assume that the n people in the poll are chosen independently and uniformly at random with replacement from the population of Pennsylvania. This is an idealized assumption, in practice certain subpopulations are more likely to be reached by polls.

Let us first perform a traditional parametric analysis of the data. Under the assumption that the people in the poll are chosen uniformly at random from the population, the probability that each of them votes for Trump is equal to the fraction of Trump voters in Pennsylvania, which we represent by a parameter θ. If the poll participants are chosen independently, then the data can be modeled as i.i.d. Bernoulli random variables (where 1 indicates a Trump vote, and 0 a Biden vote) with parameter θ. By Example 2.26, the maximum likelihood estimate of θ is the fraction of Trump voters in the poll

$$\theta_{\mathrm{ML}} = \frac{281}{581} = 0.484. \tag{6.118}$$

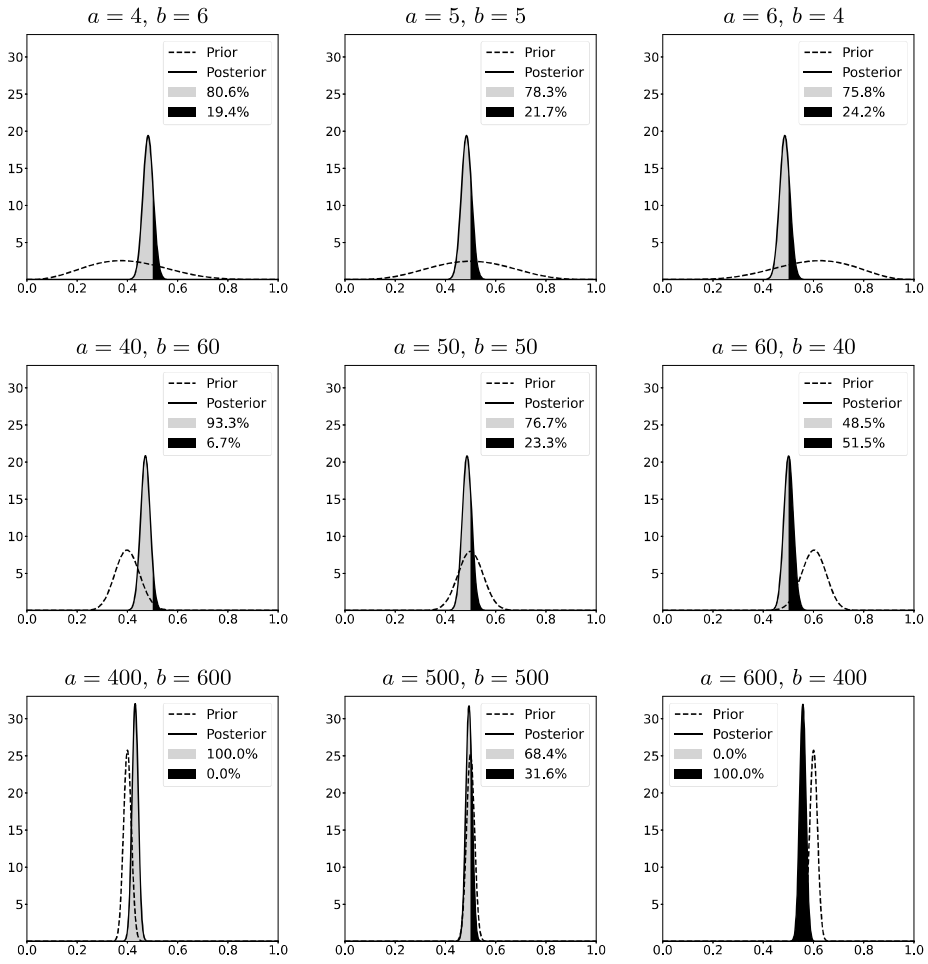

Figure 6.13 Bayesian analysis of poll data for the 2020 US presidential election. The graphs show the prior (dashed line) and posterior (solid line) pdf of the parameter $\tilde{\theta}$ in Example 6.19, which represents the fraction of Trump voters in Pennsylvania. The shaded regions under the posterior pdf indicate the probability that Biden (gray) or Trump (black) win Pennsylvania according to the model. The left column shows priors favoring Biden, the center column shows neutral priors, and the right column shows priors favoring Trump. Our confidence is encoded in the shape of the prior. In the top row, we are not very confident. In the center row, we are quite confident. In the bottom row, we are extremely confident.

The main goal of our analysis is to determine the probability that Trump or Biden wins in Pennsylvania.* Unfortunately, our traditional parametric model

*This example was written right before the 2020 election. Pennsylvania would turn out to be one of the pivotal states in the election.

does not allow us to answer this question. We are interested in the probability that θ is greater than 0.5, but this statement does not make any sense if θ is a deterministic parameter!

In order to determine the probability that Trump wins in Pennsylvania, we take a Bayesian viewpoint, where the fraction of Trump voters is modeled as a random variable $\tilde{\theta}$. We are interested in the conditional probability that $\tilde{\theta} > 0.5$ given the observed data. Under the assumption that the poll participants are chosen independently and uniformly at random with replacement from the population, the conditional distribution of the number of Trump voters $\tilde{x}$ in the poll given $\tilde{\theta} = \theta$ is binomial with parameters n and θ. If we choose the prior of $\tilde{\theta}$ to be a beta distribution with parameters a and b, by Theorem 6.18, the posterior is a beta distribution with parameters $a+281$ and $b+300$. Our probability of interest can then be computed by integrating this beta pdf between 0.5 and 1.

Figure 6.13 shows the prior and posterior pdfs, and the corresponding probability estimate, for values of the prior parameters that encode different assumptions. On the left column, the priors are favorable to Biden; in the center, they are neutral; and on the right column, they are favorable to Trump. Our confidence in the prior is encoded in the shape of the prior pdf. In the top row, we are not very confident, so the prior is very spread out. In the center row, we are quite confident, so the prior is less spread out. In the bottom row, we are extremely confident, so the prior is very concentrated.

Overall, the results of this poll seem to be quite optimistic for Biden. In most scenarios, the model predicts that he would win Pennsylvania. However, the probability that Trump wins is non-negligible for all neutral priors, and even for the low-confidence prior that favors Biden (20%). In any case, it is important to bear in mind that the probabilities output by the model are highly dependent on the prior. We must also not forget that our analysis assumes that the participants in the poll are sampled at random from the general population, which is probably not the case. In fact, the poll we are using was graded as B- by FiveThirtyEight*. For more information on modeling elections we recommend that you take a look at the methodology used by FiveThirtyEight (**?**), which is based on Bayesian modeling.

6.7.3 How Not To Predict An Election

In this section, we study the effect of independence assumptions in Bayesian probabilistic modeling. We consider the problem of predicting the outcome of the United States presidential election. In the United States, the president is elected by the Electoral College, which is formed by electors determined by the result in each state (and Washington D.C.). We can model the election result in each state as a Bernoulli random variable, which indicates whether a certain candidate wins that state or not, and estimate their joint pmf from the available data. Unfortunately, the full joint pmf has $2^{51} - 1 \geq 10^{15}$ degrees of freedom. Storing

*https://projects.fivethirtyeight.com/polls/president-general/pennsylvania/

such a joint pmf would be completely intractable, and so is estimating it from existing data. This is a manifestation of the curse of dimensionality described in Section 4.7: the dependencies in the model scale exponentially with the number of variables. Therefore, we need to make independence assumptions in order to model the data. Such independence assumptions can have a dramatic effect on the prediction produced by the model. To illustrate this, we study a simplified cartoon version of the United States presidential election.

In our cartoon presidential election, there are 51 states and each state contributes only one elector to the Electoral College. The result of the vote in state i is a Bernoulli random variable $\tilde{s}_i$ that equals one if the Republican candidate wins, and zero otherwise. The result of the whole is represented by a random variable $\tilde{e}$, which equals one if the Republican candidate wins more than 25 states and, therefore, the election:

$$\tilde{e} := \begin{cases} 1 & \text{if } \sum_{i=1}^{51} \tilde{s}_i > 25, \\ 0 & \text{otherwise.} \end{cases} \tag{6.119}$$

The Republican candidate has more support among rural voters, whereas the Democrat has more support among urban voters. Consequently, the relative turnouts of rural and urban voters in each state are crucial. We take a Bayesian perspective and model these quantities as parameters that are random variables. The rural turnout $\tilde{r}_i$, $1 \leq i \leq 51$, indicates what fraction of the voters that show up to vote on election day are rural. Conditioned on $\tilde{r}_i = r$ the probability that the Republican candidate wins state i is

$$p_{\tilde{s}_i \,|\, \tilde{r}_i}(1 \,|\, r) := 0.6r + 0.1(1 - r). \tag{6.120}$$

If the rural turnout is 100%, the Republican candidate wins the state with probability 0.6. If it is 0%, they win with probability 0.1.

We model the rural turnout in each state as a uniform random variable in $[0, 1]$. As a result, the marginal probability of the Republican candidate winning state i is

$$\begin{aligned} p_{\tilde{s}_i}(1) &= \int_{r=0}^{1} f_{\tilde{r}_i}(r) p_{\tilde{s}_i \,|\, \tilde{r}_i}(1 \,|\, r) \, \mathrm{d}r & (6.121) \\ &= \int_{r=0}^{1} (0.5r + 0.1) \, \mathrm{d}r & (6.122) \\ &= 0.35. & (6.123) \end{aligned}$$

In order to predict the result of the whole election, we need to model the joint distribution of the results in all states, which is governed by the dependence between the rural turnouts. We consider two options: (1) the rural turnouts are all mutually independent, (2) the rural turnouts are highly dependent. It turns out (no pun intended) that this is a critical decision.

Assumption 1: State Turnouts Are Independent

If the turnouts $\tilde{r}_i$ are all mutually independent for $1 \leq i \leq 51$, then so are the results in each state $\tilde{s}_i$. Under this assumption, the number of states won by the Republican candidate $\sum_{i=1}^{51} \tilde{s}_i$ is a binomial random variable with parameters $n = 51$ and $\theta = 0.35$. The Republican wins the whole election with probability

$$p_{\tilde{e}}(1) = \sum_{i=26}^{51} \binom{51}{i} 0.35^i (1 - 0.35)^{51-i} \tag{6.124}$$

$$= 0.014. \tag{6.125}$$

Under the independence assumption, it is very unlikely for the rural turnout to be simultaneously high in enough states for the Republican candidate to win the whole election.

Assumption 2: State Turnouts Are Highly Dependent

It is not unreasonable to believe that the rural turnout in different states could be dependent, especially for states that are close geographically or have similar demographics. Let us consider a situation where the rural turnout is the same in all states, i.e. $\tilde{r}_i = \tilde{r}$ for $1 \leq i \leq 51$. We still model the distribution of $\tilde{r}$ as uniform in $[0, 1]$, so the marginal distribution of the rural turnout is exactly the same as before and the probability that they win each individual state is still $p_{\tilde{s}_i}(1) = 0.35$, $1 \leq i \leq 51$.

In order to model the result of the whole election, we need to characterize the conditional *joint* distribution of $\tilde{s}_1, \ldots, \tilde{s}_{51}$ given $\tilde{r}$. To obtain a tractable model, we assume that these random variables are conditionally independent given the turnout $\tilde{r}$. This implies that the conditional distribution given $\tilde{r} = r$ of the number of states won by the Republican candidate, $\sum_{i=1}^{51} \tilde{s}_i$, is a binomial random variable with parameters $n = 51$ and $\theta = 0.6r + 0.1(1 - r)$. Consequently, the probability of the Republican candidate winning the election is

$$p_{\tilde{e}}(1) = \int_{r=0}^{1} f_{\tilde{r}}(r) p_{\tilde{e} \,|\, \tilde{r}}(1 \,|\, r) \, \mathrm{d}r \tag{6.126}$$

$$= \int_{r=0}^{1} f_{\tilde{r}}(r) \mathrm{P}\left(\sum_{i=1}^{51} \tilde{s}_i > 25 \,|\, \tilde{r} = r \right) \mathrm{d}r \tag{6.127}$$

$$= \int_{r=0}^{1} \sum_{i=26}^{51} \binom{51}{i} (0.5r + 0.1)^i (1 - (0.5r + 0.1))^{51-i} \, \mathrm{d}r \tag{6.128}$$

$$= 0.204. \tag{6.129}$$

The integral that appears in this expression does not have a closed form (although it can be computed using tabulated beta functions or Euler integrals). In practice, most realistic Bayesian models result in expressions that cannot be solved by hand. We can address this via the Monte Carlo method described in Section 1.7: simulating the election many times and then approximating the probability that

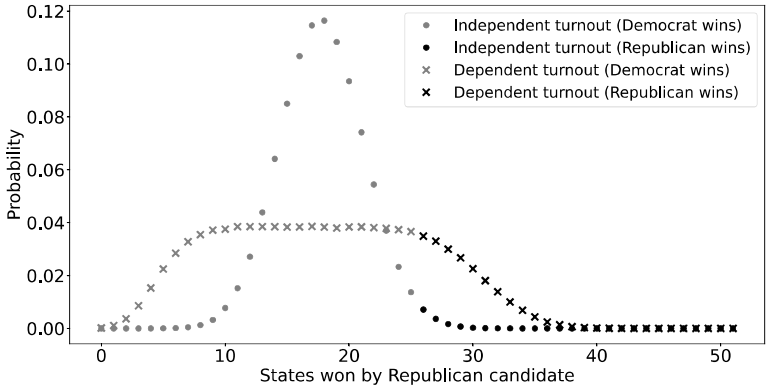

Figure 6.14 Independence assumptions can have a dramatic influence on probabilistic models. The plot shows the pmf of the states won by the Republican candidate in the idealized presidential election described in Section 6.7.3 under the assumption that the rural turnouts of different states are mutually independent (circles) or completely dependent (crosses). The probability that the Republican candidate wins more than 25 states, and therefore the whole election, correspond to the sum of the entries highlighted in black.

the Republican candidate wins using the corresponding empirical probability. In more detail, we:

1 Generate a sample r of the rural turnout by simulating a uniform distribution in the unit interval.
2 Generate the outcome in each state by sampling from independent Bernoulli random variables with parameter $0.6r + 0.1(1 - r)$.
3 Check whether the Republican candidate wins more than 25 states.

Our final estimate of $p_{\tilde{e}}(1)$ is the fraction of simulations in which the Republican candidate wins the election. This produces an accurate estimate as long as we perform enough simulations.

If the turnout is the same across the different states, the probability of the Republican candidate winning is 1 in 5, a dramatic increase from the estimate we obtained under the independent-turnout assumption (1 in 66). When rural turnout is high, it is high for all states simultaneously, which tips the election in favor of the Republican candidate. The difference between the two models is evident in the respective pmfs of the number of states won by the Republican candidate, which are shown in Figure 6.14. Ignored dependencies across states is one of the reasons why most models underestimated the probability of Trump winning in 2016. Section 9.7.3 provides an additional example of the consequences of overly optimistic independence assumptions in the context of the 2008 financial crisis.