5

Multiple Continuous Variables

Overview

In this chapter we explain how to jointly model multiple continuous quantities. Section 5.1 shows that such quantities can be represented as continuous random variables within the same probability space. These random variables can be described in terms of their joint cumulative distribution function or their joint probability density function (pdf), as discussed in Sections 5.2 and 5.3, respectively. Section 5.4 explains how to estimate the joint pdf from data, using a multidimensional generalization of kernel density estimation. Section 5.5 describes how to characterize the behavior of individual random variables in models with multiple variables. Section 5.6 defines the conditional distribution of continuous random variables given the value of other variables. Sections 5.7 and 5.8 discuss independence and conditional independence. In Section 5.9 we explain how to jointly simulate multiple continuous random variables. Finally, Section 5.10 introduces Gaussian random vectors, which are the most popular multidimensional parametric model for continuous data.

5.1 Joint Distribution Of Continuous Random Variables

Section 4.1.1 explains how to model multiple uncertain discrete quantities as random variables belonging to a common probability space. The outcome in the sample space simultaneously determines the value of all the random variables, capturing their joint behavior. In this section, we show that the same approach can be applied to model multiple continuous quantities.

As explained in Chapter 3, we describe and manipulate continuous random variables through the probability that they belong to different intervals of the real line, encoded either in their cumulative distribution function or their probability density function. It is therefore natural to describe the joint behavior of two random variables $\tilde{a}$ and $\tilde{b}$ in terms of the probability that they belong to certain intervals *at the same time*. Assuming that $\tilde{a}$ and $\tilde{b}$ are defined on the same probability space $\{\mathrm{P}, \mathcal{C}, \Omega\}$, let us consider the event that $\tilde{a} \in A$ and simultaneously $\tilde{b} \in B$, where $A, B \subset \mathbb{R}$ are two intervals, so that the two-dimensional random vector formed by $\tilde{a}$ and $\tilde{b}$ is in the rectangle $A \times B$. This event contains all outcomes ω of the sample space Ω such that $\tilde{a}(\omega)$ is in A and $\tilde{b}(\omega)$ is in B,

as illustrated in Figure 5.1. Its probability equals

$$\mathrm{P}\left(\begin{bmatrix}\tilde{a}\\ \tilde{b}\end{bmatrix} \in A \times B\right) := \mathrm{P}\left(\{\omega : \tilde{a}(\omega) \in A \text{ and } \tilde{b}(\omega) \in B\}\right) \tag{5.1}$$

$$= \mathrm{P}\left(A^{-1} \cap B^{-1}\right), \tag{5.2}$$

where

$$A^{-1} := \{\omega : \tilde{a}(\omega) \in A\}, \tag{5.3}$$

$$B^{-1} := \{\omega : \tilde{b}(\omega) \in B\}. \tag{5.4}$$

If $\tilde{a}$ and $\tilde{b}$ are continuous random variables satisfying Definition 3.1, then A^{-1} and B^{-1} belong to the collection $\mathcal{C}$ and are assigned probabilities by the probability measure P. By the properties of probability spaces (see Definitions 1.7 and 1.15), this implies that a probability must be assigned to the intersection of A^{-1} and B^{-1}, which is our event of interest. For the same reason, the probability that $\tilde{a}$ and $\tilde{b}$ belong to any pair of Borel sets (i.e. countable unions of intervals) is also well defined.

We can extend the same reasoning to describe the joint behavior of more than two continuous random variables. Let $\tilde{x}$ be a random vector containing d continuous random variables $\tilde{x}[1]$, $\tilde{x}[2]$, ..., $\tilde{x}[d]$ defined on the same probability space $(\Omega, \mathcal{C}, \mathrm{P})$. For any d intervals X_1, X_2, ..., $X_d \subseteq \mathbb{R}$, the event that $\tilde{x}$ is in the hyperrectangle $X_1 \times X_2 \times \cdots \times X_d$ can be expressed as the intersection of d events:

$$\{\omega : \tilde{x}(\omega) \in X_1 \times X_2 \times \cdots \times X_d\} = \cap_{i=1}^{d} \{\omega : \tilde{x}[i](\omega) \in X_i\}. \tag{5.5}$$

These events are all in $\mathcal{C}$ because $\tilde{x}[1]$, $\tilde{x}[2]$, ..., $\tilde{x}[d]$ are continuous random variables, so their intersection also belongs to $\mathcal{C}$ and is assigned a probability by P. Similarly, the event that $\tilde{x}$ belongs to any d-dimensional Borel set, defined as the cartesian product of d Borel sets, is also well defined.

By Theorem 3.2, we can express the probability that a continuous random variable belongs to any Borel set as a sum of the probabilities that it belongs to the individual disjoint intervals that form the set. By the same logic, we can express the probability that a d-dimensional continuous random vector belongs to a d-dimensional Borel set as the sum of the probabilities that it belongs to the individual d-dimensional hyperrectangles (cartesian products of intervals) that form the set (we can always decompose a d-dimensional Borel set in this way). The reasoning is the same as in the proof of Theorem 3.2: the events of outcomes mapping to disjoint hyperrectangles are all disjoint, so the probability of their union is equal to the sum of the individual probabilities. This means that we can describe and manipulate random vectors in terms of the probability that they belong to any d-dimensional hyperrectangle, which liberates us from having to refer to the underlying probability space. Usually, we use generalizations of the cumulative distribution function and the probability density function to keep track of these probabilities, as explained in Sections 5.2 and 5.3.

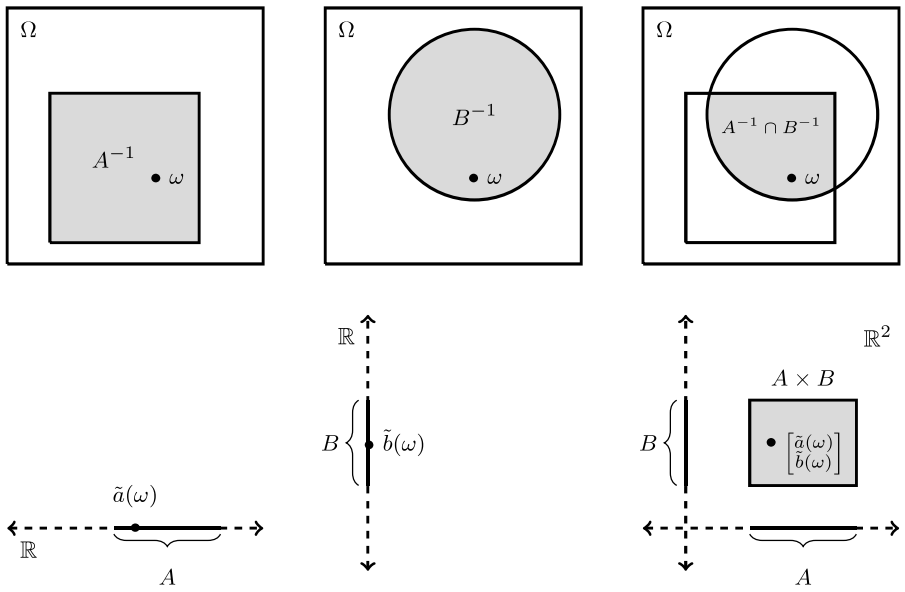

Figure 5.1 Joint distribution of continuous random variables. The continuous random variables $\tilde{a}$ and $\tilde{b}$ map outcomes in the sample space Ω to the real line $\mathbb{R}$. The left Venn diagram shows the event A^{-1} mapped by $\tilde{a}$ to the interval A, depicted below on the horizontal real line. The middle Venn diagram shows the event B^{-1} mapped by $\tilde{b}$ to the Borel set B, depicted below on the vertical real line. The right Venn diagram shows that outcomes mapped to A by $\tilde{a}$ and to B by $\tilde{b}$ are in the intersection $A^{-1} \cap B^{-1}$. These outcomes are mapped to the rectangle $A \times B$ depicted below. The probability that the vector $\left[\begin{smallmatrix}\tilde{a}\\ \tilde{b}\end{smallmatrix}\right]$ belongs to $A \times B$ therefore equals $\mathrm{P}(A^{-1} \cap B^{-1})$, represented by the area of $A^{-1} \cap B^{-1}$ in the Venn diagram.

5.2 Joint Cumulative Distribution Function

The cumulative distribution function (cdf), described in Section 3.2, encodes the probability that a random variable is less than or equal to any real number. The joint cdf of two or more random variables is a generalization of the cdf.

Definition 5.1 (Joint cumulative distribution function)**.** *Let $\tilde{a}$ and $\tilde{b}$ be random variables defined on the same probability space. The joint cumulative distribution function of $\tilde{a}$ and $\tilde{b}$ is defined as*

$$F_{\tilde{a},\tilde{b}}\,(a,b) := \mathrm{P}(\tilde{a} \leq a, \tilde{b} \leq b). \tag{5.6}$$

In words, $F_{\tilde{a},\tilde{b}}\,(a,b)$ is the probability of $\tilde{a}$ and $\tilde{b}$ being less than or equal to a and b, respectively, at the same time.

Let $\tilde{x}$ be a vector with entries equal to d continuous random variables, $\tilde{x}[i]$,

$1 \leq i \leq d$, *defined on the same probability space. The joint cdf of* $\tilde{x}$ *is*

$$F_{\tilde{x}}(x) := \mathrm{P}(\tilde{x}[1] \leq x[1], \tilde{x}[2] \leq x[2], \ldots, \tilde{x}[d] \leq x[d]). \tag{5.7}$$

In words, $F_{\tilde{x}}(x)$ *is the probability that* $\tilde{x}[i] \leq x[i]$ *for all* $i = 1, 2, \ldots, d$.

The joint cdf has similar properties to the cdf (see Lemma 3.4). For simplicity, we state them for the two-variable case, but analogous properties hold for the joint cdf of more than two random variables.

Lemma 5.2 (Properties of the joint cdf)**.** *For any random variables* $\tilde{a}$ *and* $\tilde{b}$ *defined on the same probability space, the joint cdf* $F_{\tilde{a},\tilde{b}}$ *of* $\tilde{a}$ *and* $\tilde{b}$ *satisfies*

$$\lim_{a \to -\infty} F_{\tilde{a},\tilde{b}}(a, b) = 0, \tag{5.8}$$

$$\lim_{b \to -\infty} F_{\tilde{a},\tilde{b}}(a, b) = 0, \tag{5.9}$$

$$\lim_{a \to \infty, b \to \infty} F_{\tilde{a},\tilde{b}}(a, b) = 1. \tag{5.10}$$

In addition, $F_{\tilde{a},\tilde{b}}$ *is nondecreasing,*

$$F_{\tilde{a},\tilde{b}}(a_1, b_1) \leq F_{\tilde{a},\tilde{b}}(a_2, b_2) \quad \text{if } a_2 \geq a_1, b_2 \geq b_1. \tag{5.11}$$

Proof $F_{\tilde{a},\tilde{b}}(a, b)$ is the probability that both $\tilde{a} \leq a$ and $\tilde{b} \leq b$, so if we make either of them arbitrarily small, the probability becomes zero. This can be made mathematically rigorous with the same argument we use to establish (3.16) in the proof of Lemma 3.4. Similarly, if we make a and b arbitrarily large, eventually the probability becomes one. The inequality in (5.11) holds because $\{\tilde{a} \leq a_1\} \cap \{\tilde{b} \leq b_1\}$ is a subset of $\{\tilde{a} \leq a_2\} \cap \{\tilde{b} \leq b_2\}$ so by Lemma 1.12

$$F_{\tilde{a},\tilde{b}}(a_1, b_1) = \mathrm{P}(\{\tilde{a} \leq a_1\} \cap \{\tilde{b} \leq b_1\}) \tag{5.12}$$

$$\leq \mathrm{P}(\{\tilde{a} \leq a_2\} \cap \{\tilde{b} \leq b_2\}) \tag{5.13}$$

$$= F_{\tilde{a},\tilde{b}}(a_2, b_2). \tag{5.14}$$

■

The joint cdf completely specifies the behavior of the corresponding random variables. Indeed, we can decompose any Borel set into a union of disjoint d-dimensional intervals and compute their probability by evaluating the joint cdf. In the case of two random variables $\tilde{a}$ and $\tilde{b}$, for any $a_1 < a_2$ and $b_1 < b_2$,

$$\mathrm{P}(a_1 < \tilde{a} \leq a_2, b_1 < \tilde{b} \leq b_2) \tag{5.15}$$

$$= \mathrm{P}(\{\tilde{a} \leq a_2, \tilde{b} \leq b_2\} \cap \{\tilde{a} > a_1\} \cap \{\tilde{b} > b_1\}) \tag{5.16}$$

$$= \mathrm{P}(\tilde{a} \leq a_2, \tilde{b} \leq b_2) - \mathrm{P}(\tilde{a} \leq a_1, \tilde{b} \leq b_2) - \mathrm{P}(\tilde{a} \leq a_2, \tilde{b} \leq b_1) + \mathrm{P}(\tilde{a} \leq a_1, \tilde{b} \leq b_1)$$

$$= F_{\tilde{a},\tilde{b}}(a_2, b_2) - F_{\tilde{a},\tilde{b}}(a_1, b_2) - F_{\tilde{a},\tilde{b}}(a_2, b_1) + F_{\tilde{a},\tilde{b}}(a_1, b_1), \tag{5.17}$$

This means that the joint cdf completely describes the joint behavior of the corresponding random variables. We don't have to worry about the underlying probability space. Unfortunately, it is very cumbersome to define and manipulate a joint cdf. Because of this, we tend to use probability densities instead, as described in Section 5.3. Example 5.5 explicitly computes the joint cdf for a simple

example, which illustrates how much of a pain this is, even in very simple cases. Consequently, the joint cdf is useful mostly as a mathematical tool, and is seldom estimated from data.

5.3 Joint Probability Density Function

In Section 3.3 we show that the probability density of a continuous random variable completely describes its behavior and has a very intuitive interpretation. In this section, we define a multidimensional generalization of the probability density.

Let us consider two random variables $\tilde{a}$ and $\tilde{b}$ defined on the same probability space. We are interested in defining the probability density at a certain point $\left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]$. The probability that the two-dimensional random vector formed by the two random variables belongs to a small square of area ϵ^2 that touches $\left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]$ is

$$\mathrm{P}\left(\begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} \in [a-\epsilon, a] \times [b-\epsilon, b]\right) \tag{5.18}$$

To obtain the corresponding probability density $f_{\tilde{a},\tilde{b}}(a,b)$, we divide the probability by the area of the square and take the limit $\epsilon \to 0$:

$$f_{\tilde{a},\tilde{b}}\left(a,b\right) := \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(a-\epsilon < \tilde{a} \leq a, b-\epsilon < \tilde{b} \leq b\right)}{\epsilon^2}. \tag{5.19}$$

The probability of the small square is therefore approximately equal to $\epsilon^2 f_{\tilde{a},\tilde{b}}\left(a,b\right)$. As in the one-dimensional case, the density can be obtained by differentiating the corresponding joint cdf. By (5.17) and the definition of partial derivative,

$$f_{\tilde{a},\tilde{b}}\left(a,b\right) \tag{5.20}$$

$$= \lim_{\epsilon \to 0} \frac{F_{\tilde{a},\tilde{b}}(a,b) - F_{\tilde{a},\tilde{b}}(a-\epsilon,b) - F_{\tilde{a},\tilde{b}}(a,b-\epsilon) + F_{\tilde{a},\tilde{b}}(a-\epsilon,b-\epsilon)}{\epsilon^2} \tag{5.21}$$

$$= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\lim_{\epsilon \to 0} \frac{F_{\tilde{a},\tilde{b}}(a,b) - F_{\tilde{a},\tilde{b}}(a-\epsilon,b)}{\epsilon} - \lim_{\epsilon \to 0} \frac{F_{\tilde{a},\tilde{b}}(a,b-\epsilon) - F_{\tilde{a},\tilde{b}}(a-\epsilon,b-\epsilon)}{\epsilon} \right)$$

$$= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\frac{\partial F_{\tilde{a},\tilde{b}}(a,b)}{\partial a} - \frac{\partial F_{\tilde{a},\tilde{b}}(a,b-\epsilon)}{\partial a} \right) = \frac{\partial^2 F_{\tilde{a},\tilde{b}}\left(a,b\right)}{\partial a \, \partial b}. \tag{5.22}$$

For d-dimensional random vectors, the probability density is the ratio between the probability of the random vectors belonging to a hypercube and the d-dimensional volume of the hypercube, as the volume goes to zero. As a result, it can be obtained by differentiating the joint cdf with respect to the d components of the random vector. The joint probability density function of a d-dimensional continuous random vector encodes the probability density at any point of $\mathbb{R}^d$. For it to exist, the joint cdf needs to be differentiable.

Definition 5.3 (Joint probability density function)**.** *If the joint cdf of two random variables $\tilde{a}, \tilde{b}$ defined on the same probability space is differentiable, then the*

joint pdf is

$$f_{\tilde{a},\tilde{b}}(a,b) := \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(a - \epsilon < \tilde{a} \leq a, b - \epsilon < \tilde{b} \leq b\right)}{\epsilon^2} \tag{5.23}$$

$$= \frac{\partial^2 F_{\tilde{a},\tilde{b}}\left(a,b\right)}{\partial a \partial b}. \tag{5.24}$$

If the joint cdf of a random vector $\tilde{x}$ is differentiable, then its joint pdf is

$$f_{\tilde{x}}\left(x\right)$$

$$:= \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(x[1] - \epsilon < \tilde{x}[1] \leq x[1], x[2] - \epsilon < \tilde{x}[2] \leq x[2], \ldots, x[d] - \epsilon < \tilde{x}[d] \leq x[d]\right)}{\epsilon^d}$$

$$= \frac{\partial^d F_{\tilde{x}}\left(x\right)}{\partial x[1] \, \partial x[2] \cdots \partial x[d]}. \tag{5.25}$$

The joint pdf makes it possible to compute the probability that multiple random variables, or equivalently, the entries of a random vector, belong to multidimensional Borel sets (i.e. to any countable union of hyperrectangles). In addition, the joint pdf is always nonnegative and integrates to one.

Theorem 5.4 (Properties of the joint pdf)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two continuous random variables with joint pdf $f_{\tilde{a},\tilde{b}}$. The joint pdf is nonnegative at every two-dimensional point of $\mathbb{R}^2$. In addition, for any two-dimensional Borel set $B \subseteq \mathbb{R}^2$,*

$$\mathrm{P}\left((\tilde{a},\tilde{b}) \in B\right) = \int_{(a,b) \in B} f_{\tilde{a},\tilde{b}}\left(a,b\right) \, \mathrm{d}a \, \mathrm{d}b, \tag{5.26}$$

and

$$\int_{(a,b) \in \mathbb{R}^2} f_{\tilde{a},\tilde{b}}\left(a,b\right) \, \mathrm{d}a \, \mathrm{d}b = 1. \tag{5.27}$$

Let $\tilde{x}$ be a d-dimensional continuous random vector with joint pdf $f_{\tilde{x}}$. The joint pdf is nonnegative at every d-dimensional point of $\mathbb{R}^d$. In addition, for any d-dimensional Borel set $B \subseteq \mathbb{R}^d$,

$$\mathrm{P}\left(\tilde{x} \in B\right) = \int_{x \in B} f_{\tilde{x}}\left(x\right) \, \mathrm{d}x \tag{5.28}$$

and

$$\int_{x \in \mathbb{R}^d} f_{\tilde{x}}\left(x\right) \, \mathrm{d}x = 1. \tag{5.29}$$

Conversely, any function $f : \mathbb{R}^d \to \mathbb{R}$ that is nonnegative and integrates to one over $\mathbb{R}^d$ can be interpreted as the joint pdf of a random vector $\tilde{x} : \Omega \to \mathbb{R}^d$ for some sample space Ω.

Proof The joint pdf is nonnegative because the joint cdf is nondecreasing with respect to all of its entries, as established in Lemma 5.2.

Integrating the joint pdf over a Borel set B yields the probability that the corresponding random vector belongs to B by the same reasoning as in one dimension (see Section 3.3). We can partition B into hypersquares of volume ϵ^d (in

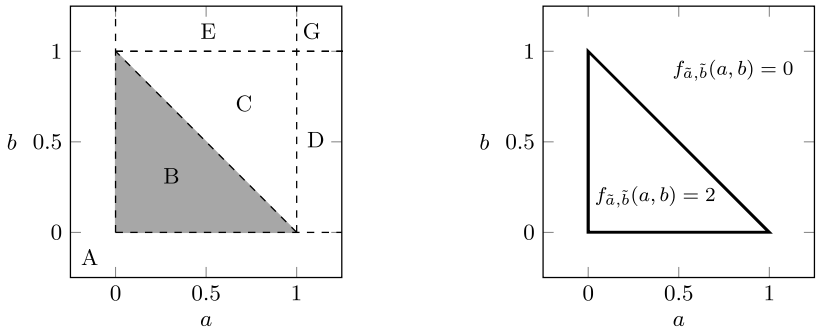

Figure 5.2 Triangle lake in Example 5.5. The left plot shows the lake shaded in gray. The different regions marked A to F are used to define the joint cdf. The right plot shows the corresponding joint pdf.

two dimensions, squares of area ϵ^2). The probability that the random vector is in B is the sum of the probabilities that it is in any of the hyperrectangles. As $\epsilon \to 0$, the probability is given by the value of the joint pdf in the hyperrectangle multiplied by ϵ^d. The sum of these probabilities is a Riemann sum that converges to the integral of the joint pdf over B. This establishes (5.26) and (5.28). A proof based on the joint cdf, generalizing the proof of Theorem 3.16, is also possible.

The integral of the joint pdf over the whole space $\mathbb{R}^d$ must equal one because every point in the original sample space of the probability space associated to the random vector must map to some point in $\mathbb{R}^d$. To prove that any nonnegative function f with this property is a valid joint pdf, we can use it to construct a probability measure on the sample space $\mathbb{R}^d$, choosing the collection of Borel sets as the associated collection of events, and the probabilities obtained from f as the probability measure. ■

In summary, the joint pdf completely characterizes the joint behavior of the corresponding random variables, as illustrated in the following example.

Example 5.5 (Triangle lake)**.** A biologist is tracking an otter that lives in the triangular lake depicted in Figure 5.2. She decides to model the location of the otter probabilistically, as a two-dimensional random vector with entries $\tilde{a}$ and $\tilde{b}$. The otter does not seem to prefer any specific region of the lake, so the biologist decides to model the location as being uniformly distributed, meaning that the joint pdf of $\tilde{a}$ and $\tilde{b}$ is constant,

$$f_{\tilde{a},\tilde{b}}\,(a,b) = \begin{cases} c & \text{if } (a,b) \in \text{Lake}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.30}$$

To find the normalizing constant c we use the fact that the joint pdf must integrate

to one:

$$\int_{a=-\infty}^{\infty}\int_{b=-\infty}^{\infty} c \, \mathrm{d}a \, \mathrm{d}b = \int_{b=0}^{1}\int_{a=0}^{1-b} c \, \mathrm{d}a \, \mathrm{d}b \tag{5.31}$$

$$= c\int_{b=0}^{1} (1-b) \, \mathrm{d}b \tag{5.32}$$

$$= \frac{c}{2} = 1, \tag{5.33}$$

so $c = 2$. Figure 5.2 shows the joint pdf.

To compute the probability that the otter is in any subset of the lake, we integrate the joint pdf over the corresponding subset. For example,

$$\mathrm{P}(\{\tilde{a} \geq 0.6, \tilde{b} \leq 0.2\}) = \int_{b=0}^{0.2}\int_{a=0.6}^{1-b} 2 \, \mathrm{d}a \, \mathrm{d}b \tag{5.34}$$

$$= \int_{b=0}^{0.2} 2\,(0.4-b) \, \mathrm{d}b \tag{5.35}$$

$$= 2(0.08 - 0.02) \tag{5.36}$$

$$= 0.12. \tag{5.37}$$

We now compute the joint cdf of $\tilde{a}$ and $\tilde{b}$. $F_{\tilde{a},\tilde{b}}\,(a,b)$ represents the probability that the otter is southwest of the point (a,b). Computing the joint cdf requires dividing the range into the sets shown in Figure 5.2 and integrating the joint pdf. If $(a,b) \in A$, then $F_{\tilde{a},\tilde{b}}\,(a,b) = 0$ because $\mathrm{P}\left(\{\tilde{a} \leq a\} \cap \{\tilde{b} \leq b\}\right) = 0$. If $(a,b) \in B$,

$$F_{\tilde{a},\tilde{b}}\,(a,b) = \int_{u=0}^{b}\int_{v=0}^{a} 2 \, \mathrm{d}v \, \mathrm{d}u = 2ab. \tag{5.38}$$

If $(a,b) \in C$,

$$F_{\tilde{a},\tilde{b}}\,(a,b) = \int_{u=0}^{1-a}\int_{v=0}^{a} 2 \, \mathrm{d}v \, \mathrm{d}u + \int_{u=1-a}^{b}\int_{v=0}^{1-u} 2 \, \mathrm{d}v \, \mathrm{d}u = 2a + 2b - b^2 - a^2 - 1.$$

Setting $a := 1$ in this expression, yields the value of the joint cdf at $(a,b) \in D$, since

$$F_{\tilde{a},\tilde{b}}\,(a,b) = \mathrm{P}\left(\tilde{a} \leq a, \tilde{b} \leq b\right) = \mathrm{P}\left(\tilde{a} \leq 1, \tilde{b} \leq b\right) = 2b - b^2. \tag{5.39}$$

Exchanging the roles of a and b, we obtain $F_{\tilde{a},\tilde{b}}\,(a,b) = 2a - a^2$ for $(a,b) \in E$ by the same reasoning. Finally, for $(a,b) \in G$ $F_{\tilde{a},\tilde{b}}\,(a,b) = 1$ because $\mathrm{P}(\tilde{a} \leq a, \tilde{b} \leq b) = 1$. Putting everything together,

$$F_{\tilde{a},\tilde{b}}\,(a,b) = \begin{cases} 0 & \text{if } a < 0 \text{ or } b < 0, \\ 2ab, & \text{if } a \geq 0, b \geq 0, a + b \leq 1, \\ 2a + 2b - b^2 - a^2 - 1, & \text{if } a \leq 1, b \leq 1, a + b \geq 1, \\ 2b - b^2, & \text{if } a \geq 1, 0 \leq b \leq 1, \\ 2a - a^2, & \text{if } 0 \leq a \leq 1, b \geq 1, \\ 1, & \text{if } a \geq 1, b \geq 1. \end{cases} \tag{5.40}$$

This was rather painful! The joint pdf is usually a much more convenient way of describing a distribution than the joint cdf.

5.4 Multidimensional Density Estimation

In this section we show how to estimate the joint pdf of continuous random variables using a multidimensional extension of kernel density estimation (see Section 3.5.2). The idea is the same as in the univariate case: the density estimate is a superposition of shifted copies of a kernel function centered at the data points. The only difference is that the kernel function is multidimensional.

Definition 5.6 (Multidimensional kernel density estimation)**.** *Consider a d-dimensional real-valued dataset $X := \{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathbb{R}^d$ for $1 \leq i \leq n$. The corresponding kernel density estimate at any point $a \in \mathbb{R}^d$ is*

$$f_{X,h}(a) := \frac{1}{n\,h^d} \sum_{i=1}^{n} K\left(\frac{a - x_i}{h}\right), \tag{5.41}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel function centered at the origin that satisfies

$$K(a) \geq 0 \quad \text{for all } a \in \mathbb{R}^d, \tag{5.42}$$

$$\int_{\mathbb{R}^d} K(a)\, \mathrm{d}a = 1. \tag{5.43}$$

The parameter $h \geq 0$ governs the spread of the kernel function, while ensuring that it integrates to one. Generalizations where h is a matrix that induces different spreads in different directions are also possible.

A popular choice for the kernel is the Gaussian function

$$K(a) := \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{||a||_2^2}{2}\right), \tag{5.44}$$

which is smooth and decays rapidly away from its center. Figure 5.3 shows a simple example, where we apply kernel density estimation with a Gaussian kernel to three two-dimensional data points. The density estimate is shown as a three-dimensional surface, as well as a contour plot.

As in the case of one-dimensional kernel density estimation, the bandwidth h has a strong influence on the density estimate. Increasing h dilates the kernel, so that the density estimate takes into account more points. If the bandwidth is very small, individual samples have a large influence on the density estimate. This makes it possible to reproduce irregular shapes more easily, but may also overfit spurious fluctuations in the observed data, especially if we don't have many observations. Increasing the bandwidth smooths out such fluctuations and yields more stable estimates. However, a value of h that is too large results in over-smoothing, and can eliminate meaningful structure from the estimate. The

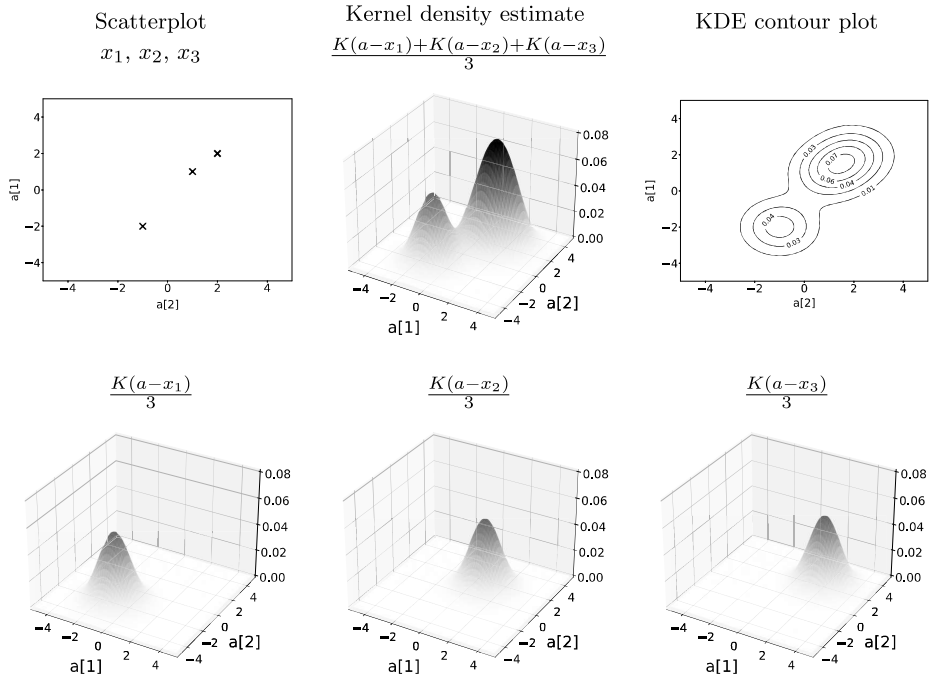

Figure 5.3 Kernel density estimation in 2D. We apply kernel density estimation to approximate the joint probability density corresponding to three data points x_1, x_2, x_3 shown in the upper left corner. The density is estimated by summing shifted copies of a Gaussian kernel with bandwidth equal to one ($h := 1$ in Definition 5.6) centered at each data point. Each of these components is shown in the bottom row. The result is shown in the graph at the center of the top row. On the right, we show the corresponding contour plot.

tradeoff is illustrated in Figure 5.4, where we show the contour plots of density estimates obtained using different bandwidths. The density corresponds to temperatures measured hourly at weather stations in Manhattan (Kansas) and Versailles (Kentucky) in 2015, extracted from Dataset 9.

It is important to point out that kernel density estimation is useful for modeling the joint pdf of a small number of variables. When we consider models with many variables, we run into the curse of dimensionality. Section 4.7 describes this phenomenon for discrete random variables: as the number of variables increases, the number of entries in the joint pmf explodes exponentially. Modeling continuous random variables is even more challenging, because each of them can take values in a continuous set. As a result, estimating the joint pdf using a nonparametric method such as kernel density estimation is usually intractable, unless the number of variables is small. Instead, we often resort to parametric models such as Gaussian random vectors, described in Section 5.10.

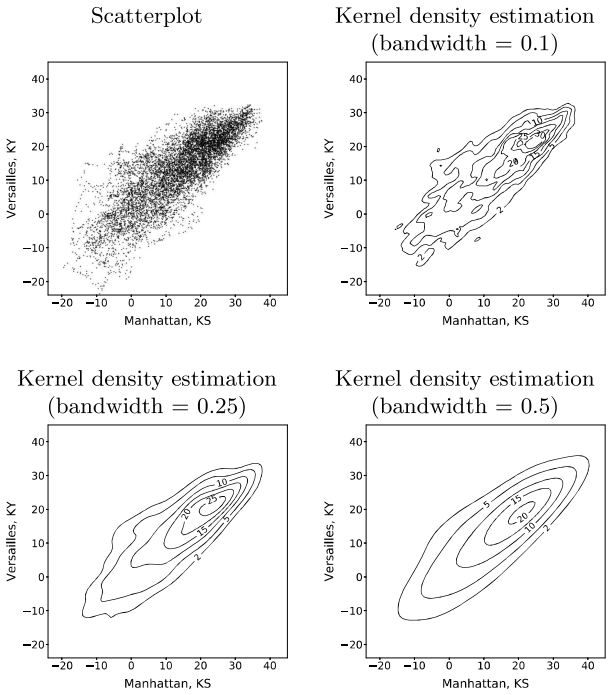

Figure 5.4 Temperature in Kansas and Kentucky. The top left graph shows a scatterplot of the data, which are temperatures measured hourly at weather stations in Manhattan (Kansas) and Versailles (Kentucky) in 2015. The remaining images show the corresponding kernel density estimates obtained using Gaussian kernels with different bandwidths.

5.5 Marginal Distributions

In probabilistic models consisting of multiple random variables, it is often useful to isolate the individual behavior of a single variable. In the case of discrete random variables, this is achieved by summing the joint pmf over the rest of the variables (see Section 4.2). In the case of continuous random variables, we can obtain the pdf of a single variable, which we call the *marginal* pdf, in a similar way: by integrating the joint pdf with respect to the remaining variables.

Theorem 5.7 (Marginal pdf)**.** *Let $\tilde{a}$ and $\tilde{b}$ be continuous random variables with joint pdf $f_{\tilde{a},\tilde{b}}$. The marginal pdf of $\tilde{a}$ is*

$$f_{\tilde{a}}(a) = \int_{b=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}(a,b)\, \mathrm{d}b. \tag{5.45}$$

Let $\tilde{x}$ be a d-dimensional random vector. The marginal pdf of $\tilde{x}[i]$ is

$$f_{\tilde{x}[i]}(a) \tag{5.46}$$
$$= \int_{b_1 \in \mathbb{R}} \cdots \int_{b_{i-1} \in \mathbb{R}} \int_{b_{i+1} \in \mathbb{R}} \cdots \int_{b_d \in \mathbb{R}} f_{\tilde{x}}\left(b[1], \ldots, b[i-1], a, b[i+1], \ldots, b[d]\right) \mathrm{d}b_1 \ldots \mathrm{d}b_{i-1} \, \mathrm{d}b_{i+1} \ldots \mathrm{d}b_d.$$

Proof We prove the bivariate case, the vector case follows by the same argument. By (5.26)

$$F_{\tilde{a}}(a) = \mathrm{P}(\tilde{a} \leq a) \tag{5.47}$$
$$= \int_{u=-\infty}^{a} \int_{b=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}(u, b) \, \mathrm{d}b \, \mathrm{d}u. \tag{5.48}$$

Taking the derivative with respect to a completes the proof. ■

Alternatively, we can estimate the marginal cdf of a random variable from the joint cdf by computing the limit when the remaining variables tend to infinity.

Lemma 5.8 (Marginal cumulative distribution function)**.** *Let $\tilde{a}$ and $\tilde{b}$ be continuous random variables with joint cdf $F_{\tilde{a},\tilde{b}}$. The marginal cdf of $\tilde{a}$ is*

$$F_{\tilde{a}}(a) = \lim_{b \to \infty} F_{\tilde{a},\tilde{b}}(a, b). \tag{5.49}$$

Proof When $b \to \infty$, the limit of $F_{\tilde{a},\tilde{b}}(a, b)$ is the probability of $\tilde{a}$ being smaller than a, which is precisely the marginal cdf of $\tilde{a}$. More formally,

$$\lim_{b \to \infty} F_{\tilde{a},\tilde{b}}(a, b) = \lim_{n \to \infty} \mathrm{P}\left(\cup_{i=1}^{n} \{\tilde{a} \leq a, \tilde{b} \leq i\}\right) \tag{5.50}$$
$$= \mathrm{P}\left(\lim_{n \to \infty} \{\tilde{a} \leq a, \tilde{b} \leq i\}\right) \tag{5.51}$$
$$= \mathrm{P}(\tilde{a} \leq a) \tag{5.52}$$
$$= F_{\tilde{a}}(a). \tag{5.53}$$

■

If we are interested in computing the joint pdf of several entries in a random vector, instead of just one, the marginalization process is essentially the same. Notation gets a bit complicated, so let us consider an example with $d := 4$. To compute the marginal joint pdf of the first and fourth entry of random vector $\tilde{x}$, we integrate the joint pdf with respect to the second and third entries,

$$f_{\tilde{x}[1],\tilde{x}[4]}(a, d) = \int_{b=-\infty}^{\infty} \int_{c=-\infty}^{\infty} f_{\tilde{x}}(a, b, c, d) \, \mathrm{d}b \, \mathrm{d}c. \tag{5.54}$$

Example 5.9 (Triangle lake: Marginal distribution)**.** The biologist wants to compute the probability density of the horizontal coordinate of the otter. The

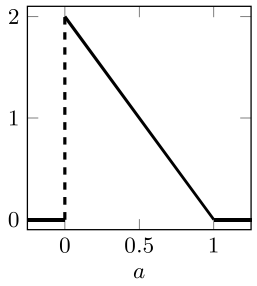

Figure 5.5 Marginal pdf for the triangle lake. The plot shows the marginal pdf $f_{\tilde{a}}$ of the horizontal coordinate $\tilde{a}$ of the otter in Example 5.5.

marginal pdf of $\tilde{a}$ is obtained by integrating over $\tilde{b}$, which yields

$$f_{\tilde{a}}(a) = \int_{b=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}(a,b)\ \mathrm{d}b \tag{5.55}$$

$$= \int_{b=0}^{1-a} 2\ \mathrm{d}b \tag{5.56}$$

$$= 2(1-a) \tag{5.57}$$

for $0 \leq a \leq 1$ and zero otherwise. The pdf is shown in Figure 5.5.

Figure 5.6 shows estimates of the marginal pdfs of the temperature in Manhattan (Kansas) and Versailles (Kentucky) in 2015. To estimate the marginal densities, we can integrate the joint pdf obtained via multidimensional KDE, as explained in Section 5.4, but there is a simpler alternative: applying one-dimensional KDE to the temperature values at each station separately.

5.6 Conditional Distributions

Conditional distributions allow us to update our uncertainty about certain variables in a model when other variables have been observed. For example, we may want to estimate the *conditional* pdf of a random variable $\tilde{b}$ when another random variable $\tilde{a}$, defined on the same probability space, equals a. The conditional density of $\tilde{b}$ at a point b should equal

$$\lim_{\epsilon_1 \to 0} \frac{\mathrm{P}\left(b - \epsilon_1 < \tilde{b} \leq b \,|\, \tilde{a} = a\right)}{\epsilon_1}. \tag{5.58}$$

Unfortunately, this conditional probability is not well defined when $\tilde{a}$ is continuous. The problem is that the probability of the event $\tilde{a} = a$ is zero (see Section 3.1). However, we can condition on the event that $\tilde{a}$ belongs to a small neighborhood of length ϵ_2 touching a. The conditional density given that event

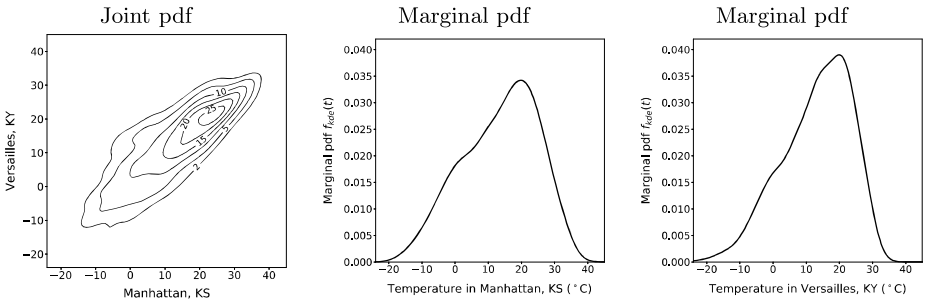

Figure 5.6 Marginal densities for temperature in Kansas and Kentucky. The left graph shows a contour plot of the estimated joint pdf of the temperature in Manhattan (Kansas) and Versailles (Kentucky) in 2015. The center and right graphs show the estimated marginal densities of the temperatures in Manhattan and Versailles respectively, obtained by applying one-dimensional KDE to the corresponding temperatures.

is

$$f_{\tilde{b}\,|\,a-\epsilon_2<\tilde{a}\leq a}(b) := \lim_{\epsilon_1\to 0} \frac{\mathrm{P}\left(b-\epsilon_1<\tilde{b}\leq b\,|\,a-\epsilon_2<\tilde{a}\leq a\right)}{\epsilon_1}. \tag{5.59}$$

To obtain a conditional density given the event $\tilde{a}=a$, we take the limit when $\epsilon_2\to 0$, so that the neighborhood collapses to a. The resulting density can be expressed as the ratio between the joint pdf of $\tilde{a}$ and $\tilde{b}$ and the marginal pdf of $\tilde{a}$. To simplify the derivation, we set ϵ_1 and ϵ_2 to equal a single value denoted by ϵ:

$$\begin{aligned}
f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) &= \lim_{\epsilon\to 0} f_{\tilde{b}\,|\,a-\epsilon<\tilde{a}\leq a}(b) && (5.60)\\
&= \lim_{\epsilon\to 0} \frac{\mathrm{P}\left(b-\epsilon<\tilde{b}\leq b\,|\,a-\epsilon<\tilde{a}\leq a\right)}{\epsilon} && (5.61)\\
&= \lim_{\epsilon\to 0} \frac{1}{\epsilon}\frac{\mathrm{P}\left(b-\epsilon<\tilde{b}\leq b, a-\epsilon<\tilde{a}\leq a\right)}{\mathrm{P}\left(a-\epsilon<\tilde{a}\leq a\right)} && (5.62)\\
&= \frac{\lim_{\epsilon\to 0}\frac{\mathrm{P}\left(b-\epsilon<\tilde{b}\leq b, a-\epsilon<\tilde{a}\leq a\right)}{\epsilon^2}}{\lim_{\epsilon\to 0}\frac{\mathrm{P}\left(a-\epsilon<\tilde{a}\leq a\right)}{\epsilon}} && (5.63)\\
&= \frac{f_{\tilde{a},\tilde{b}}\left(a,b\right)}{f_{\tilde{a}}\left(a\right)}, && (5.64)
\end{aligned}$$

as long as $f_{\tilde{a}}\left(a\right)>0$. Inspired by this calculation, we define the conditional pdf of a continuous random variable given another continuous random variable as the ratio of the joint pdf and the marginal pdf of the variable that we are conditioning on.

Definition 5.10 (Conditional pdf)**.** *Let $\tilde{a},\tilde{b}$ be random variables with joint pdf*

$f_{\tilde{a},\tilde{b}}$ defined on the same probability space. The conditional pdf of $\tilde{b}$ given $\tilde{a}$ is

$$f_{\tilde{b}\,|\,\tilde{a}}\,(b\,|\,a) := \frac{f_{\tilde{a},\tilde{b}}\,(a,b)}{f_{\tilde{a}}\,(a)} \qquad \text{if } f_{\tilde{a}}\,(a) > 0 \tag{5.65}$$

and is undefined otherwise.

Let $\tilde{x}$ be a d-dimensional random vector. The conditional pdf of $\tilde{x}[i]$ given $\tilde{x}[j] = a_j$ for $j \neq i$ is

$$f_{\tilde{x}[i]\,|\,\tilde{x}[1],\ldots,\tilde{x}[i-1],\tilde{x}[i+1],\ldots,\tilde{x}[d]}(b\,|\,a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_d) \tag{5.66}$$

$$= \frac{f_{\tilde{x}}(a_1,\ldots,a_{i-1},b,a_{i+1},\ldots,a_d)}{f_{\tilde{x}[1],\ldots,\tilde{x}[i-1],\tilde{x}[i+1],\ldots,\tilde{x}[d]}(a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_d)}, \tag{5.67}$$

if the denominator is nonzero.

We can compute the conditional cdf of $\tilde{b}$ given $\tilde{a}$ using the conditional pdf,

$$F_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) := \int_{u=-\infty}^{b} f_{\tilde{b}\,|\,\tilde{a}}\,(u\,|\,a)\;\mathrm{d}u. \tag{5.68}$$

The conditional pdf is a valid pdf, since it is nonnegative and it integrates to one.

Lemma 5.11. *Let $\tilde{a}, \tilde{b}$ be random variables with joint pdf $f_{\tilde{a},\tilde{b}}$. The conditional pdf of $\tilde{b}$ given $\tilde{a}$ integrates to one. Let $\tilde{x}$ be a d-dimensional random vector. The conditional pdf of $\tilde{x}[i]$ given $\tilde{x}[j] = x[j]$, for $j \neq i$, also integrates to one.*

Proof We prove the bivariate case, the proof for the vector case is the same. For any a such that $f_{\tilde{a}}\,(a) > 0$,

$$\int_{b=-\infty}^{\infty} f_{\tilde{b}\,|\,\tilde{a}}\,(b\,|\,a)\;\mathrm{d}b = \frac{\int_{b=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}\,(a,b)\;\mathrm{d}b}{f_{\tilde{a}}\,(a)} \tag{5.69}$$

$$= \frac{f_{\tilde{a}}\,(a)}{f_{\tilde{a}}\,(a)} = 1. \tag{5.70}$$

■

We can also define the conditional pdf of some of the entries in a random vector given other entries. For an example with $d = 4$, the conditional pdf of the second and third entries given the first and fourth entries is

$$f_{\tilde{x}[2],\tilde{x}[3]\,|\,\tilde{x}[1],\tilde{x}[4]}(b,c\,|\,a,d) = \frac{f_{\tilde{x}}(a,b,c,d)}{f_{\tilde{x}[1],\tilde{x}[4]}(a,d)}. \tag{5.71}$$

All such conditional pdfs integrate to one, as long as we integrate them over the appropriate variables. For example, the conditional pdf $f_{\tilde{x}[2],\tilde{x}[3]\,|\,\tilde{x}[1],\tilde{x}[4]}$ in (5.71) integrates to one with respect to $x[2]$ and $x[3]$, but it does not make sense to integrate it with respect to $x[1]$ and $x[4]$.

An immediate consequence of Definition 5.10 is the chain rule for continuous random variables.

Theorem 5.12 (Chain rule for continuous random variables)**.** *For any continuous random variables $\tilde{a}, \tilde{b}$ with joint pdf $f_{\tilde{a},\tilde{b}}$,*

$$f_{\tilde{a},\tilde{b}}(a,b) = f_{\tilde{a}}(a)\, f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) \tag{5.72}$$

$$= f_{\tilde{b}}(b)\, f_{\tilde{a}\,|\,\tilde{b}}(a\,|\,b)\,. \tag{5.73}$$

Similarly, the joint pdf of any d-dimensional random vector $\tilde{x}$ with joint pdf $f_{\tilde{x}}$ can be decomposed into

$$f_{\tilde{x}}(x) = f_{\tilde{x}[1]}(x[1]) \prod_{i=2}^{n} f_{\tilde{x}[i]\,|\,\tilde{x}[1],\ldots,\tilde{x}[i-1]}(x[i]\,|\,x[1],\ldots,x[i-1])\,. \tag{5.74}$$

The order of indices in the random vector is completely arbitrary (any order works).

Example 5.13 (Triangle lake: Conditional distribution)**.** The biologist wants to compute the probability density of the vertical coordinate of the otter when she knows the horizontal coordinate. The conditional pdf is

$$f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) = \frac{f_{\tilde{a},\tilde{b}}(a,b)}{f_{\tilde{a}}(a)} \tag{5.75}$$

$$= \frac{1}{1-a} \qquad 0 \leq b \leq 1-a, \tag{5.76}$$

where $f_{\tilde{a}}$ is calculated in Example 5.9. For fixed $\tilde{a} = a$, the density of probability for $\tilde{b}$ is constant between 0 and $1-a$, as illustrated in Figure 5.7. On the upper right we show the marginal pdf of $f_{\tilde{a}}$. Each value is computed by integrating the joint pdf along a vertical line. In particular, $f_{\tilde{a}}(0.75)$ is obtained by integrating along the dashed line shown on the joint-pdf plot. The lower left plot shows the value of the joint pdf on that line, i.e. $f_{\tilde{a},\tilde{b}}(0.75, b)$. This is not a conditional pdf; its integral equals $f_{\tilde{a}}(0.75)$ instead of one. When we compute the conditional pdf, we are effectively normalizing this quantity, dividing it by $f_{\tilde{a}}(0.75)$, as shown on the lower right.

Example 5.14 (Conditional distribution of temperature)**.** Figure 5.8 shows an estimate of the conditional pdf of the temperature in Versailles (Kentucky) given the temperature in Manhattan (Kansas), obtained using hourly measurements from 2015. To obtain the estimate, we divide the joint pdf approximated via 2D kernel density estimation (see Figure 5.4) by the marginal pdf of the temperature in Manhattan approximated via 1D kernel density estimation (see Figure 5.6). The results reveal that there is a strong dependence between the temperatures at the two locations: each conditional pdf is approximately centered at the corresponding temperature in Manhattan.

Figure 5.9 shows estimates of the conditional joint pdf of the temperatures in Versailles (Kentucky) and Corvallis (Oregon) in 2015 given different temperatures in Manhattan (Kansas). We compute the conditional pdf by applying the

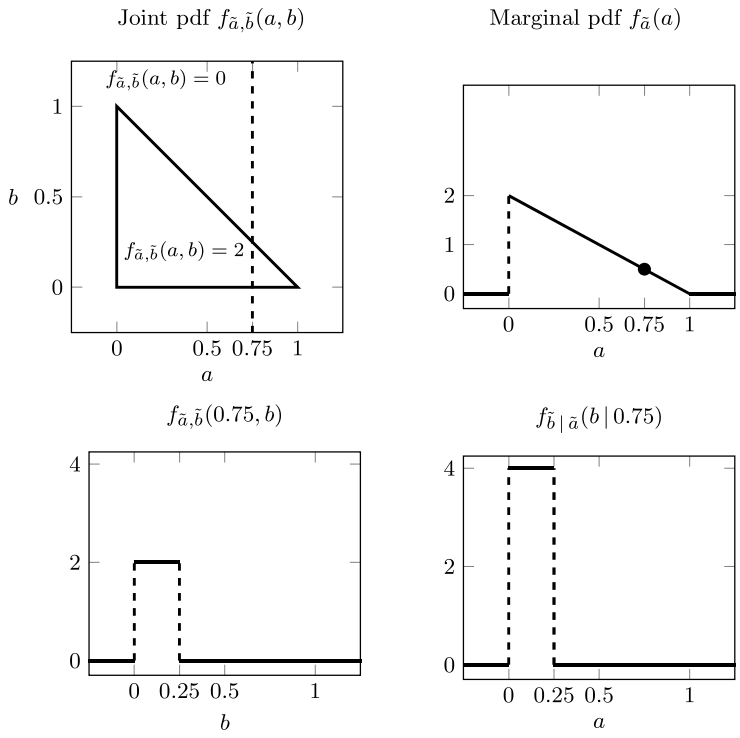

Figure 5.7 Conditional pdf in the triangle lake. The upper left diagram shows the joint pdf from Example 5.5 and a dashed line corresponding to the values for which $\tilde{a} = 0.75$. The upper right plot shows the marginal pdf of $f_{\tilde{a}}$. The value $f_{\tilde{a}}(0.75)$ is indicated by a circular marker. This value is computed by integrating along the dashed line shown on the joint-pdf plot. The lower left plot shows the value of the joint pdf on that line, i.e. $f_{\tilde{a},\tilde{b}}(0.75, b)$. The lower right plot shows the conditional pdf of $\tilde{b}$ given $\tilde{a} = 0.75$, obtained by normalizing $f_{\tilde{a},\tilde{b}}(a, 0.75)$ so that it integrates to one. To this end, we divide by $f_{\tilde{a}}(0.75)$, which is precisely equal to the integral of $f_{\tilde{a},\tilde{b}}(0.75, b)$.

definition. Let $\tilde{v}$, $\tilde{c}$ and $\tilde{m}$ be random variables representing the temperature in Versailles, Corvallis and Manhattan, respectively. We have

$$f_{\tilde{v},\tilde{c}\,|\,\tilde{m}}(v, c\,|\,t) = \frac{f_{\tilde{v},\tilde{c},\tilde{m}}(v, c, t)}{f_{\tilde{m}}(t)}, \tag{5.77}$$

where the joint pdf in the numerator and the marginal pdf in the denominator can both be approximated via kernel density estimation.

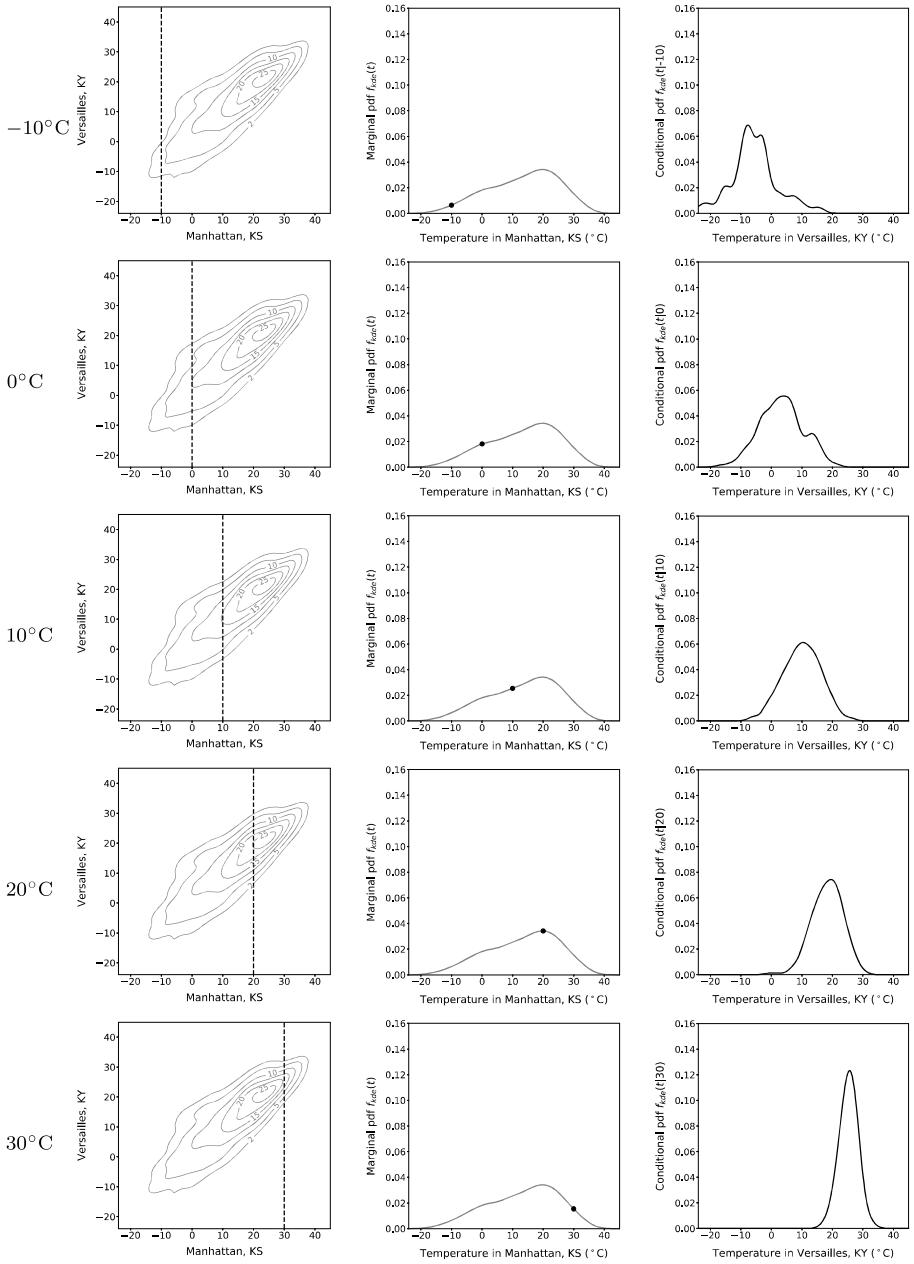

Figure 5.8 Temperature in Kentucky, conditioned on temperature in Kansas. The left column shows the contour plot of the estimated joint pdf of the temperature in Manhattan (Kansas) and Versailles (Kentucky) in 2015. In each row, we condition on a different temperature in Manhattan. The conditional density of the temperature in Versailles is obtained by dividing the joint pdf at that value (dashed line in first column) by the corresponding value of the marginal pdf of Manhattan in the second column (indicated by a black marker). The result is shown in the third column.

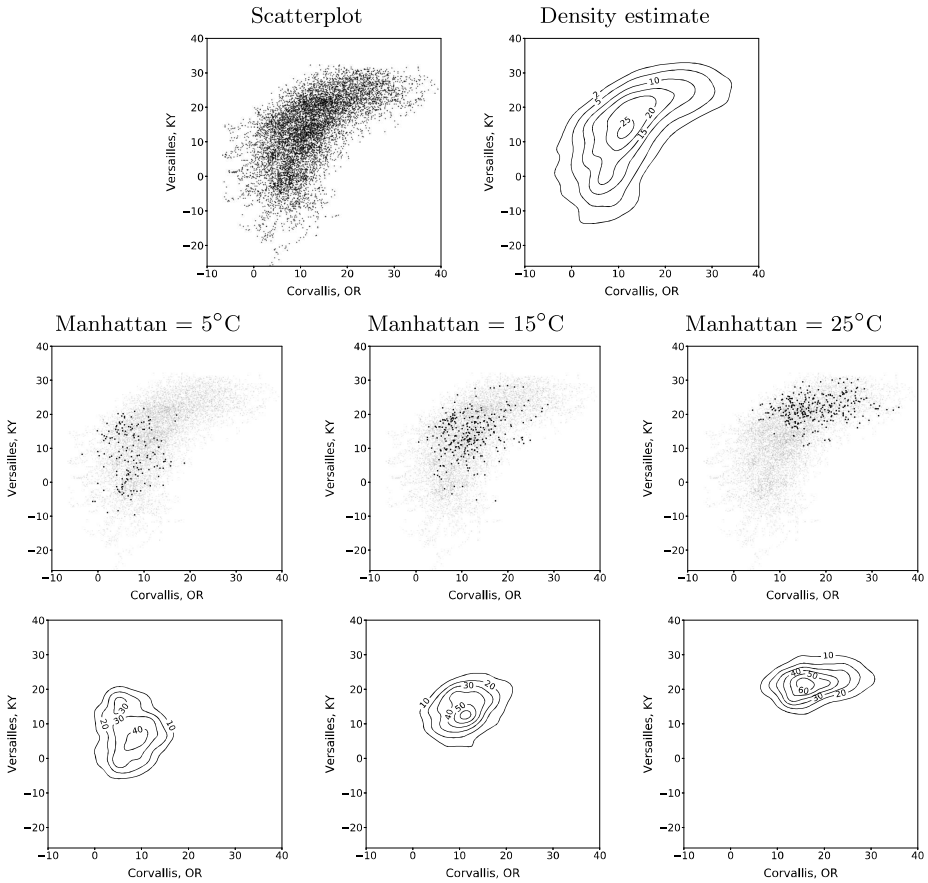

Figure 5.9 Temperature in Kentucky and Oregon, conditioned on the temperature in Kansas. The top column shows the scatterplot (left) and corresponding contour plot of the estimated joint pdf (right) of the temperature in Versailles (Kentucky) and Corvallis (Oregon) in 2015. The second row shows the same scatterplot, where we highlight the data points for which the temperature in Manhattan (Kansas) is in $[t - 0.5, t + 0.5]$ for different values of t. The third row shows the estimated conditional joint pdf of the temperature in Versailles and Corvallis given different temperatures in Manhattan. The estimate is obtained by dividing the joint pdf of the three stations by the marginal pdf of Manhattan (both calculated via kernel density estimation).

5.7 Independence

When knowledge about a random variable $\tilde{a}$ does not affect our uncertainty about another random variable $\tilde{b}$, we say that $\tilde{a}$ and $\tilde{b}$ are *independent*. This can be expressed in terms of the conditional distributions of the random variables. If the

variables are continuous, then we should have

$$P(\tilde{a} \in S \,|\, \tilde{b} = b) = \mathrm{P}(\tilde{a} \in S) \tag{5.78}$$

for any Borel set S and any b. This is the case if and only if

$$\begin{aligned} F_{\tilde{a}\,|\,\tilde{b}}(a \,|\, b) &= P(\tilde{a} \leq a \,|\, \tilde{b} = b) && (5.79)\\ &= P(\tilde{a} \leq a) && (5.80)\\ &= F_{\tilde{a}}(a) && (5.81) \end{aligned}$$

for any $a, b \in \mathbb{R}$. If the cdfs are differentiable, this is equivalent to $f_{\tilde{a}\,|\,\tilde{b}}(a \,|\, b) = f_{\tilde{a}}(a)$, or

$$f_{\tilde{a},\tilde{b}}(a, b) = f_{\tilde{a}}(a) f_{\tilde{b}}(b). \tag{5.82}$$

Definition 5.15 (Independent random variables)**.** *Two continuous random variables $\tilde{a}$ and $\tilde{b}$ defined on the same probability space are independent if and only if*

$$f_{\tilde{a},\tilde{b}}\,(a, b) = f_{\tilde{a}}\,(a)\, f_{\tilde{b}}\,(b)\,, \quad \text{for all } (a, b) \in \mathbb{R}^2. \tag{5.83}$$

The definition can be extended to multiple random variables and random vectors. In order to ensure that all possible conditional probabilities are the same as the marginal probabilities, we require the joint pdfs to factorize completely.

Definition 5.16 (Independent random vectors)**.** *The d entries $\tilde{x}[1]$, $\tilde{x}[2]$, ..., $\tilde{x}[d]$ in a continuous random vector $\tilde{x}$ are independent if and only if*

$$f_{\tilde{x}}\,(x) = \prod_{i=1}^{d} f_{\tilde{x}[i]}\,(x[i])\,, \quad \text{for all } x \in \mathbb{R}^d. \tag{5.84}$$

Recall that when we discussed parametric models, we defined the likelihood $\mathcal{L}_X$ of a dataset X, as the product of the individual likelihoods. This is equivalent to defining the likelihood as the joint pdf of the data (interpreted as a function of the parameters) under the assumption that all variables are independent and have the same marginal distribution. We say that such random variables are independent and identically distributed, analogously to Definition 4.19.

Definition 5.17 (Independent identically distributed random variables)**.** *Let $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ be continuous random variables belonging to the same probability space. The random variables are identically distributed if their marginal pdfs are the same*

$$f_{\tilde{a}_1} = f_{\tilde{a}_2} = \cdots = f_{\tilde{a}_n} = f_{\tilde{a}} \tag{5.85}$$

for some pdf $f_{\tilde{a}}$. They are independent and identically distributed (i.i.d.) if their joint pdf is equal to the product of the marginal pdfs

$$f_{\tilde{a}_1,\tilde{a}_2,\ldots,\tilde{a}_n}\,(a_1, a_2, \ldots, a_n) = \prod_{i=1}^{d} f_{\tilde{a}}\,(a_i)\,, \tag{5.86}$$

for all possible values of a_1, a_2, ..., a_n.

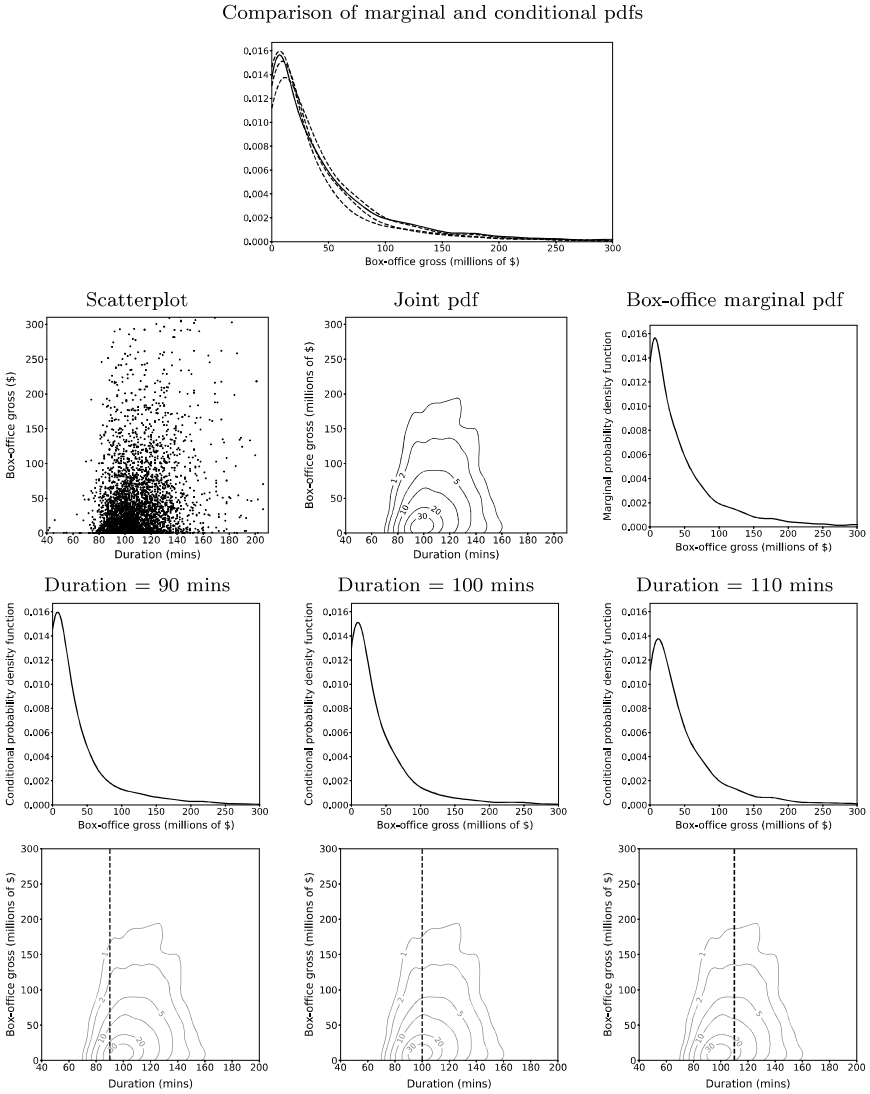

Figure 5.10 Movie duration and box-office earnings are approximately independent. The top plot compares the marginal and conditional densities of the box-office earnings of movies, showing that they are quite similar. The second row shows the scatterplot of box-office earnings and duration (left), the contour plot of the corresponding joint pdf obtained via KDE (center), and the marginal pdf of the earnings (right). The third row shows the estimated conditional pdf of the earnings conditioned on different durations. The bottom row shows the different slices of the joint pdf (indicated by a dashed line in the contour plot) that are used to estimate the conditional pdf.

Figure 5.10 shows an example of two quantities that are approximately independent: the duration of a movie and its box-office revenues, extracted from Dataset 11. The conditional pdf (estimated via kernel density estimation) of the box-office earnings given different durations (90, 100, and 110 minutes) are very similar, indicating that the duration of the movie alone does not strongly determine the distribution of the box-office earnings (at least over that range of durations). Intuitively, independence implies that no matter how we *slice* the joint pdf of the two random variables to compute the conditional density, we always end up with the same density (see the bottom row of the figure). The example illustrates the difficulty of establishing that two quantities are completely independent in practice. Since we never have access to infinite data, we can only approximate the densities, so they will never be exactly the same. In this case, it looks like longer movies may have somewhat larger earnings, but the effect is small.

5.8 Conditional Independence

Figure 5.12 shows that the temperatures in Versailles (Kentucky) and Corvallis (Oregon) are not independent. The distribution of temperatures in Versailles changes dramatically depending on the temperature in Corvallis. This is evident in the shape of the estimated joint pdf between the two temperatures (third row in Figure 5.12). However, when we condition on the temperature in Manhattan (Kansas) then there is much less dependence between the temperatures in Versailles and Corvallis, as shown in Figure 5.13. The conditional pdf of the temperature in Versailles looks very similar whether we condition on Manhattan, or on both Manhattan and Corvallis. It seems that the temperature at Corvallis provides very little information *as long as we also know the temperature in Manhattan*. The reason is the geographic location of the weather stations, shown in Figure 5.11: Manhattan is situated approximately between Corvallis and Versailles.

If our uncertainty about a random variable $\tilde{a}$ does not change when another random variable $\tilde{b}$ is revealed, *as long as the value of a third random variable* $\tilde{c}$ *is known*, then $\tilde{a}$ and $\tilde{b}$ are conditionally independent given $\tilde{c}$. In our example, Corvallis and Versailles are not conditionally independent given Manhattan, because they do have some conditional dependence (the conditional pdfs in Figure 5.13 are not exactly the same). For conditional independence to hold, we require

$$P(\tilde{a} \in S \,|\, \tilde{b} = b, \tilde{c} = c) = \mathrm{P}(\tilde{a} \in S \,|\, \tilde{c} = c) \tag{5.87}$$

for any Borel set S and any possible c, which holds if and only if

$$F_{\tilde{a} \,|\, \tilde{b},\tilde{c}}(a \,|\, b, c) = F_{\tilde{a} \,|\, \tilde{c}}(a \,|\, c) \tag{5.88}$$

for all $a, b \in \mathbb{R}$. If the cdfs are differentiable, this is equivalent to $f_{\tilde{a} \,|\, \tilde{b},\tilde{c}}(a \,|\, b, c) = f_{\tilde{a} \,|\, \tilde{c}}(a \,|\, c)$ or

$$f_{\tilde{a},\tilde{b} \,|\, \tilde{c}}(a, b \,|\, c) = f_{\tilde{a} \,|\, \tilde{c}}(a \,|\, c) f_{\tilde{b} \,|\, \tilde{c}}(b \,|\, c). \tag{5.89}$$

Figure 5.11 Weather stations in Oregon, Kansas and Kentucky. Location of the weather stations of Corvallis (Oregon), Manhattan (Kansas) and Versailles (Kentucky). The geographic configuration of the weather stations explains why the temperature in Versailles and Corvallis are approximately conditionally independent given the temperature in Manhattan (see Figure 5.13).

Definition 5.18 (Conditionally independent random variables)**.** *Two continuous random variables $\tilde{a}$ and $\tilde{b}$ defined on the same probability space are conditionally independent given a random variable $\tilde{c}$ if and only if*

$$f_{\tilde{a},\tilde{b}\,|\,\tilde{c}}\,(a,b\,|\,c) = f_{\tilde{a}\,|\,\tilde{c}}\,(a\,|\,c)\,f_{\tilde{b}\,|\,\tilde{c}}\,(b\,|\,c)\,, \quad \textit{for all } a,b,c \in \mathbb{R}. \tag{5.90}$$

The definition can be extended to multiple random variables or random vectors, conditioned on multiple random variables. Notation can get a bit complicated, but the main idea is the same as independence: all conditional distributions must equal the marginals, which implies that the joint pmf and pdfs factorize into the product of the marginals.

Definition 5.19 (Conditionally independent random vectors)**.** *The d_1 entries $\tilde{x}[1]$, $\tilde{x}[2]$, ..., $\tilde{x}[d_1]$ in a continuous random vector $\tilde{x}$ are conditionally independent given a d_2-dimensional random vector $\tilde{y}$ if and only if*

$$f_{\tilde{x}\,|\,\tilde{y}}\,(x\,|\,y) = \prod_{i=1}^{d} f_{\tilde{x}[i]\,|\,\tilde{y}}\,(x[i]\,|\,y)\,, \quad \textit{for all } x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}. \tag{5.91}$$

As discussed in Section 1.6, independence does *not* imply conditional independence or vice versa.

5.9 Jointly Simulating Multiple Random Variables

In Section 3.8 we explain how to simulate an arbitrary distribution via inverse-transform sampling. Here, we explain how to simulate the joint distribution of multiple random variables.

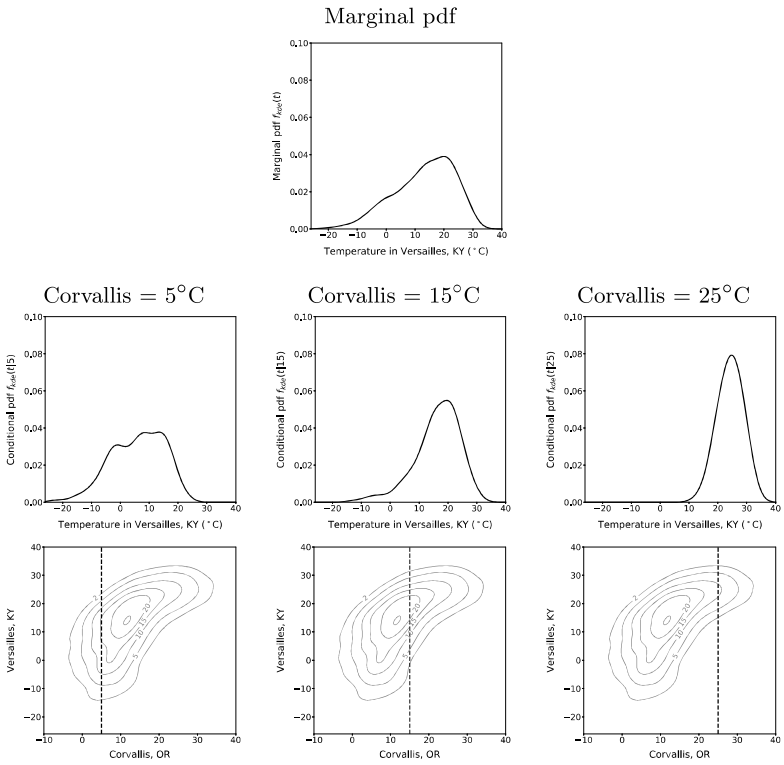

Figure 5.12 The temperatures in Kentucky and Oregon are not independent. The top row shows the marginal pdf of the temperature in Versailles (Kentucky) in 2015 estimated via KDE. The second row shows estimates of the conditional pdf given different temperatures in Corvallis (Oregon). The marginal and conditional pdfs are clearly very different. This is reflected in the shape of the estimated joint pdf of both temperatures, shown in the third row. The conditional pdfs correspond to normalized *slices* of the joint pdf, indicated by the dashed lines.

Let us first focus on the case of two random variables $\tilde{a}$ and $\tilde{b}$. We can apply inverse-transform sampling to simulate $\tilde{a}$ or $\tilde{b}$ separately, but this would not produce samples from the joint distribution. We illustrate this for the joint pdf from Example 5.5 in Figure 5.14. The center scatterplot shows the result of separately simulating the horizontal and vertical coordinates according to their marginal distributions. Each coordinate has the right distribution, *but the joint distribution is wrong!* There are many samples outside of the triangle, and they are clearly not uniformly distributed within the triangle.

The problem with our approach is that the conditional distribution of the horizontal coordinate given the vertical coordinate is incorrect (and vice versa). This can be addressed by simulating the random variables sequentially. First, we

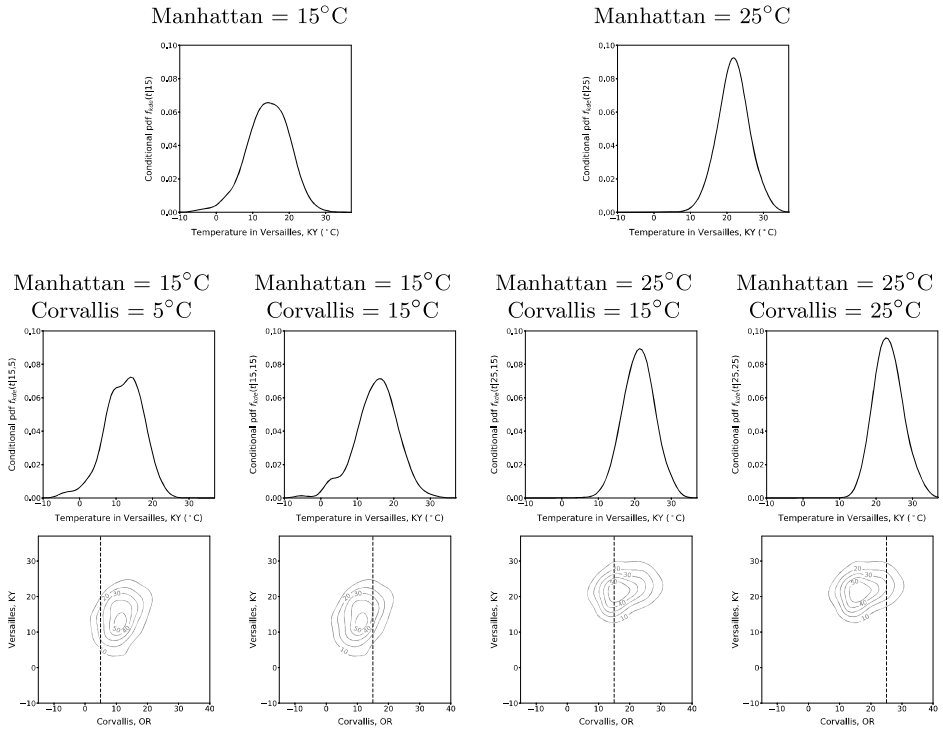

Figure 5.13 The temperatures in Kentucky and Oregon are almost conditionally independent given the temperature in Kansas. The top row shows the conditional pdf of the temperature in Versailles (Kentucky) given different temperatures in Manhattan (Kansas). The second row shows the conditional pdf when we also condition on the temperature in Corvallis (Oregon). The difference between conditioning on Corvallis or not is rather small (compare to the difference when we do not condition on Manhattan, see Figure 5.12). This is reflected in the shape of the estimated conditional joint pdf of both temperatures given the temperature in Manhattan, shown in the third row. The conditional pdfs correspond to normalized *slices* of the conditional joint pdf indicated by the dashed lines.

generate a sample a from the marginal distribution of $\tilde{a}$. Then, we generate a sample b from the *conditional distribution of* $\tilde{b}$ *given* $\tilde{a} = a$. This can be achieved by applying inverse-transform sampling to the conditional cdf of $\tilde{b}$ given $\tilde{a} = a$. The resulting pair of values (a, b) have the correct joint distribution, as illustrated by the graph on the right of Figure 5.14.

Example 5.20 (Triangle lake: Sampling from the joint distribution)**.** In order to generate samples from the joint pdf in Example 5.5, we first generate a sample from $\tilde{a}$ via inverse-transform sampling. Integrating the marginal pdf of $\tilde{a}$ derived in Example 5.9 yields the marginal cdf of $\tilde{a}$, which equals $F_{\tilde{a}}(a) = 2a - a^2$ for

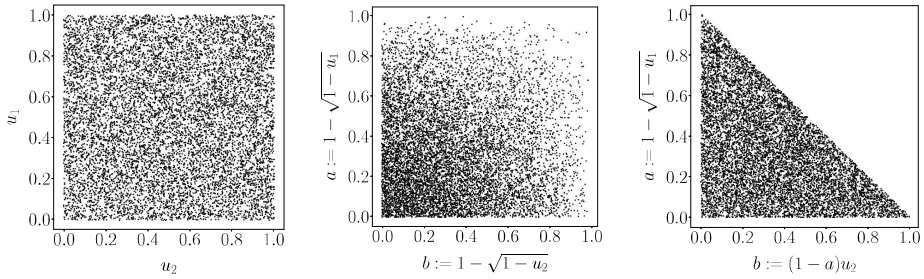

Figure 5.14 Simulating a triangular joint pdf. The left plot shows 10,000 i.i.d. samples where the horizontal and vertical coordinates are independent and uniformly distributed in $[0, 1]$. The central plot shows the effect of transforming the vertical and horizontal coordinates via inverse-transform sampling, so that they are distributed according to the marginal distributions of $\tilde{a}$ and $\tilde{b}$ in Example 5.20. This does not produce the right joint distribution. The right plot shows the effect of transforming the vertical coordinate to simulate the marginal distribution of $\tilde{a}$, and the horizontal coordinate to simulate the conditional distribution of $\tilde{b}$ given the corresponding value of the vertical coordinate. This does simulate the correct joint distribution of $\tilde{a}$ and $\tilde{b}$.

$a \in [0, 1]$. The inverse is $F_{\tilde{a}}^{-1}(u) = 1 - \sqrt{1-u}$. We generate a sample from the marginal distribution of $\tilde{a}$ by plugging a sample u_1 from the uniform distribution in $[0, 1]$ into $F_{\tilde{a}}^{-1}$:

$$a_{\text{samp}} := 1 - \sqrt{1 - u_1}. \tag{5.92}$$

We can apply the same approach to simulate the marginal distribution of $\tilde{b}$, which by symmetry is exactly the same as the marginal distribution of $\tilde{a}$. Using another sample u_2 from the uniform distribution in $[0, 1]$ this yields:

$$b_{\text{wrong}} := 1 - \sqrt{1 - u_2}. \tag{5.93}$$

As illustrated by the graph at the center of Figure 5.14, this doesn't work. The problem is that we are simulating the two random variables *independently*. To capture the dependence between the random variables, we instead sample from the conditional distribution of $\tilde{b}$ given $\tilde{a}$. Consider a fixed sample a_{samp} of $\tilde{a}$. As derived in Example 5.13, the conditional pdf of $\tilde{b}$ given $\tilde{a} = a_{\text{samp}}$ equals $(1 - a_{\text{samp}})^{-1}$ in $[0, 1 - a_{\text{samp}}]$. The conditional cdf is therefore $F_{\tilde{b} \,|\, \tilde{a}}(b \,|\, a_{\text{samp}}) = (1 - a_{\text{samp}})^{-1} b$. The inverse equals $F_{\tilde{b} \,|\, \tilde{a}}^{-1}(u \,|\, a_{\text{samp}}) = (1 - a_{\text{samp}})u$. We obtain a sample from the conditional distribution by plugging the uniform sample u_2 into $F_{\tilde{b} \,|\, \tilde{a}}^{-1}(\cdot \,|\, a_{\text{samp}})$:

$$b_{\text{samp}} := (1 - a_{\text{samp}})u_2. \tag{5.94}$$

The pair of values $(a_{\text{samp}}, b_{\text{samp}})$ correctly capture the dependence between $\tilde{a}$ and $\tilde{b}$, as shown in the graph on the right of Figure 5.14.

5.10 Gaussian Random Vectors

As mentioned in Section 5.4, nonparametric density estimation is often intractable unless the number of variables is small. This motivates the use of parametric models to estimate the joint pdf of continuous random vectors. Gaussian random vectors are a multidimensional generalization of Gaussian random variables, described in Section 3.6.2. They are perhaps the most popular parametric multidimensional models for continuous data. In Section 5.10.1 we derive the parametric distribution of Gaussian random vectors. In Section 5.10.2 we study the marginal and conditional distributions of the entries of these vectors, establishing that they are all Gaussian. In Section 5.10.3 we explain how to fit the multidimensional Gaussian distribution to data via maximum-likelihood estimation.

5.10.1 Definition

In this section, we explain how to derive a parametric model for multiple variables based on the Gaussian distribution. The most straightforward approach is to model each variable separately. Let $\tilde{x}$ denote a d-dimensional random vector. If we model the ith entry as a Gaussian random variable with mean μ_i and standard deviation σ_i for $1 \leq i \leq d$, and we assume that the entries are all independent, then the joint pdf of the random vector equals,

$$f_{\tilde{x}}(x) = \prod_{i=1}^{d} f_{\tilde{x}[i]}\left(\tilde{x}[i]\right) \tag{5.95}$$

$$= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\left(x[i]-\mu_i\right)^2}{2\sigma_i^2}\right) \tag{5.96}$$

$$= \frac{1}{(2\pi)^{\frac{d}{2}} \prod_{i=1}^{d} \sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{d} \frac{\left(x[i]-\mu_i\right)^2}{\sigma_i^2}\right). \tag{5.97}$$

This model can easily be fit to data by applying the maximum-likelihood estimator derived in Theorem 3.36 to each variable separately. Figure 5.15 shows an example with just two variables. Unfortunately, our model is pretty terrible. The problem is that it cannot capture the dependence between the variables. This is evident in the shape of the contour surfaces of the parametric density, which are surfaces on which the density is constant:

$$\left\{x \in \mathbb{R}^d : f_{\tilde{x}}(x) = c\right\} = \left\{x \in \mathbb{R}^d : \sum_{i=1}^{d} \frac{\left(x[i]-\mu_i\right)^2}{\sigma_i^2} = c'\right\}, \tag{5.98}$$

where $c \geq 0$ is a fixed constant and $c' = -2\log\left(c\,(2\pi)^{\frac{d}{2}} \prod_{i=1}^{d} \sigma_i\right)$. The contour surfaces are concentric ellipsoids with axes that are aligned with the coordinate axes. In the case of Figure 5.15, they are 2D ellipses with axes that are constrained to be horizontal and vertical. The standard deviation σ_i determines the length of the ith axis and the means $\mu_1, \ldots, \mu_2$ determine the center of the ellipsoids.

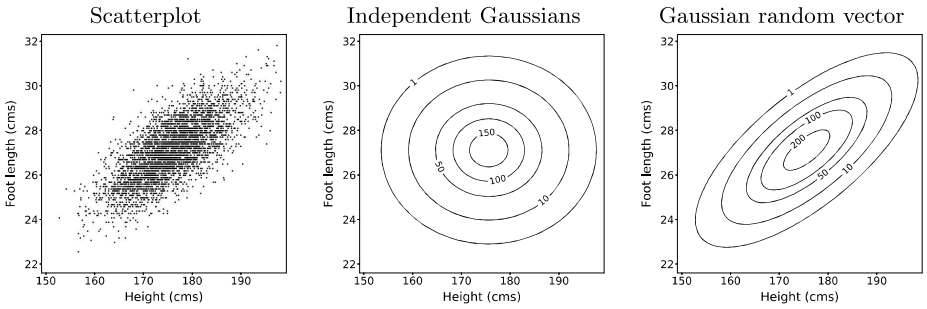

Figure 5.15 The Gaussian parametric model in multiple dimensions. The left image shows a scatterplot of the height and foot length of 4,082 men in the United States army, extracted from Dataset 5. The middle image shows the contour lines of a Gaussian parametric model where each entry is modeled independently. The model is not able to capture the dependence between height and foot length (taller people have longer feet). The third column shows the density of a parametric Gaussian model that is able to capture this dependence (the model is fit via maximum likelihood as described in Section 5.10.3).

Looking at the data in Figure 5.15, it is clear that fixing the ellipsoid axes to lie along the coordinate axes is too constraining: it is impossible to achieve a good fit just by stretching the ellipses horizontally or vertically.

In order to provide more flexibility to our parametric model, we allow for *rotations* of the density with respect to the coordinate axes. This can be achieved by incorporating additional parameters: d orthonormal vectors $u_1, u_2, \ldots, u_d$. The contour surfaces of an ellipsoid with axes aligned with these vectors equal

$$\left\{ x \in \mathbb{R}^d : \sum_{i=1}^{d} \frac{\left(u_i^T \left(x - \mu\right)\right)^2}{\sigma_i^2} = c' \right\}, \tag{5.99}$$

for some constant c'. Here μ is the vector of mean parameters, such that $\mu[i] = \mu_i$, $1 \leq i \leq d$. We can reformulate the equation for the ellipsoid as a quadratic function:

$$\sum_{i=1}^{d} \frac{\left(u_i^T \left(x - \mu\right)\right)^2}{\sigma_i^2} = \left(x - \mu\right)^T U \Lambda^{-1} U^T \left(x - \mu\right) \tag{5.100}$$

$$= \left(x - \mu\right)^T \Sigma^{-1} \left(x - \mu\right) \tag{5.101}$$

where

$$U := \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}, \qquad \Lambda := \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}, \qquad \Sigma := U \Lambda U^T.$$

By the spectral theorem (Theorem 11.19), any positive definite symmetric ma-

trix Σ has an eigendecomposition of the form $U\Lambda U^T$ where U is orthogonal and Λ is diagonal. This means that we can use the matrix parameter Σ to encode the variance parameters $\sigma_1, \ldots, \sigma_d$ and the axes parameters $u_1, \ldots, u_d$. In order to obtain our desired flexible model, we replace the exponent in equation (5.97) by the quadratic function in (5.101):

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right), \tag{5.102}$$

Including the determinant $|\Sigma|$ in the denominator of (5.102) ensures that the expression integrates to one. Note that Σ needs to be positive definite for $f_{\tilde{x}}$ to be a valid joint pdf. If it is not, then there is a vector v for which $v^T\Sigma v < 0$. In that direction, the density would explode to infinity (take $x := \mu + \alpha v$ where $\alpha \to \infty$). We call Σ the covariance-matrix parameter, because it is in fact the covariance matrix of the random vector (see Theorem 11.10). Notice that the joint pdf of our model no longer factorizes into a product of the marginals, so the entries of the random vector $\tilde{x}$ are not independent. Figure 5.15 shows that this can be very important to fit multiple variables effectively.

To summarize, we have derived a model that has ellipsoidal contours rotated with respect to the coordinate axes. Random vectors that have this parametric distribution are known as Gaussian random vectors.

Definition 5.21 (Gaussian random vector)**.** *A Gaussian random vector $\tilde{x}$ of dimension d is a random vector with joint pdf*

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right), \tag{5.103}$$

where $|\Sigma|$ denotes the determinant of Σ. The joint pdf is parametrized by the mean vector $\mu \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Σ must be invertible, symmetric and positive definite (all its eigenvalues must be positive).

The following example shows how to derive the ellipsoidal contours of a Gaussian joint pdf. This requires reverse engineering our derivation of the covariance-matrix parameter Σ using its eigendecomposition.

Example 5.22 (Two-dimensional Gaussian)**.** We consider a two-dimensional Gaussian random vector where μ is the zero vector and

$$\Sigma = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix}. \tag{5.104}$$

Since μ is zero, the contour lines of the density correspond to the set of points where $x^T\Sigma^{-1}x$ is constant. The eigenvalues of Σ are $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, and the corresponding eigenvectors equal

$$u_1 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}, \quad u_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}. \tag{5.105}$$

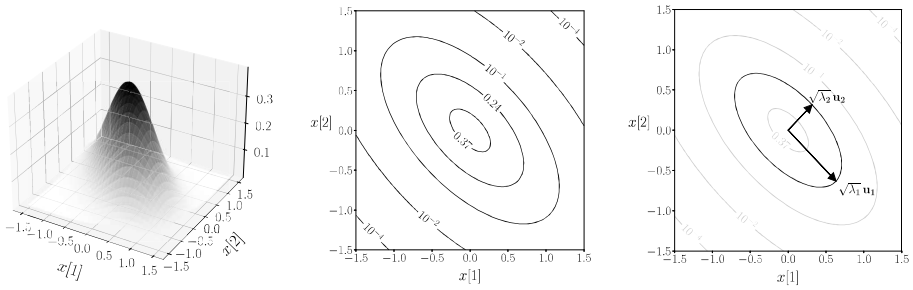

Figure 5.16 Probability density of a Gaussian vector. The left image shows the joint pdf of the two-dimensional Gaussian random vector defined in Example 5.22. The middle plot shows the corresponding contour plot. The axes align with the eigenvectors of the covariance matrix. Their lengths are proportional to the square root of the corresponding eigenvalues, as shown on the right image for a specific contour line.

The contour lines are therefore ellipses with equation

$$x^T \Sigma^{-1} x = \frac{(u_1^T x)^2}{\lambda_1} + \frac{(u_2^T x)^2}{\lambda_2} = c', \tag{5.106}$$

for some constant c'. Figure 5.16 shows a 3D plot of the density and its ellipsoidal contour lines with axes aligned with u_1 and u_2. The rightmost plot in the figure shows the ellipse where the density equals 0.24.

5.10.2 Marginal And Conditional Distributions

In this section we study the marginal and conditional distributions of Gaussian random vectors. We begin with a simple example in two dimensions.

Example 5.23 (Two-dimensional Gaussian random vector)**.** Let us consider a two-dimensional Gaussian random vector $(\tilde{a}, \tilde{b})$ with zero mean and covariance matrix

$$\Sigma := \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{5.107}$$

where ρ is a parameter, which we call the *correlation coefficient* of the vector. The correlation coefficient of two random variables captures their linear dependence, as described in detail in Chapter 8. For now, we interpret it as a parameter determining the shape of the distribution. In order for Σ to be positive definite, we need $-1 < \rho < 1$ (you can check this by computing the eigenvalues). The inverse of the covariance matrix is

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}. \tag{5.108}$$

The joint pdf of the random vector equals

$$f_{\tilde{a},\tilde{b}}(a,b) := \frac{1}{\sqrt{(2\pi)^2 \left|\Sigma\right|}} \exp\left(-\frac{1}{2}\begin{bmatrix} a \\ b \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} a \\ b \end{bmatrix}\right) \tag{5.109}$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{a^2 - 2\rho ab + b^2}{2(1-\rho^2)}\right) \tag{5.110}$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{(1-\rho^2)a^2 + (b-\rho a)^2}{2(1-\rho^2)}\right) \tag{5.111}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(b-\rho a)^2}{2(1-\rho^2)}\right). \tag{5.112}$$

These algebraic manipulations are often referred to as *completing the square*. The term

$$\frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(b-\rho a)^2}{2(1-\rho^2)}\right) \tag{5.113}$$

is a Gaussian pdf with mean ρa and variance $1-\rho^2$ (see Definition 3.31), which means that it integrates to one with respect to b. The marginal pdf of $\tilde{a}$ therefore equals

$$f_{\tilde{a}}(a) = \int_{b=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}(a,b)\, \mathrm{d}b \tag{5.114}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) \int_{b=-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(b-\rho a)^2}{2(1-\rho^2)}\right) \mathrm{d}b \tag{5.115}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right). \tag{5.116}$$

The marginal distribution of $\tilde{a}$ is Gaussian with zero mean and unit variance. The exact same argument switching a and b implies that the marginal distribution of $\tilde{b}$ is also Gaussian with zero mean and unit variance.

By the definition of conditional pdf,

$$f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) = \frac{f_{\tilde{a},\tilde{b}}(a,b)}{f_{\tilde{a}}(a)} \tag{5.117}$$

$$= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(b-\rho a)^2}{2(1-\rho^2)}\right). \tag{5.118}$$

The conditional distribution of $\tilde{b}$ given $\tilde{a} = a$ is Gaussian with mean ρa and variance $1-\rho^2$. The correlation coefficient completely determines the dependence between $\tilde{a}$ and $\tilde{b}$. When it is zero, they are independent because $f_{\tilde{b}\,|\,\tilde{a}}$ is equal to the marginal pdf of $\tilde{b}$. When ρ is close to 1 and $\tilde{a} = a$, $f_{\tilde{b}\,|\,\tilde{a}}$ is closely concentrated (the variance tends to zero) near a (the mean approaches ρa). When ρ is close to -1, the density concentrates around $-a$ instead. Figure 5.17 shows the joint pdf for different values of ρ, as well as the corresponding conditional pdf.

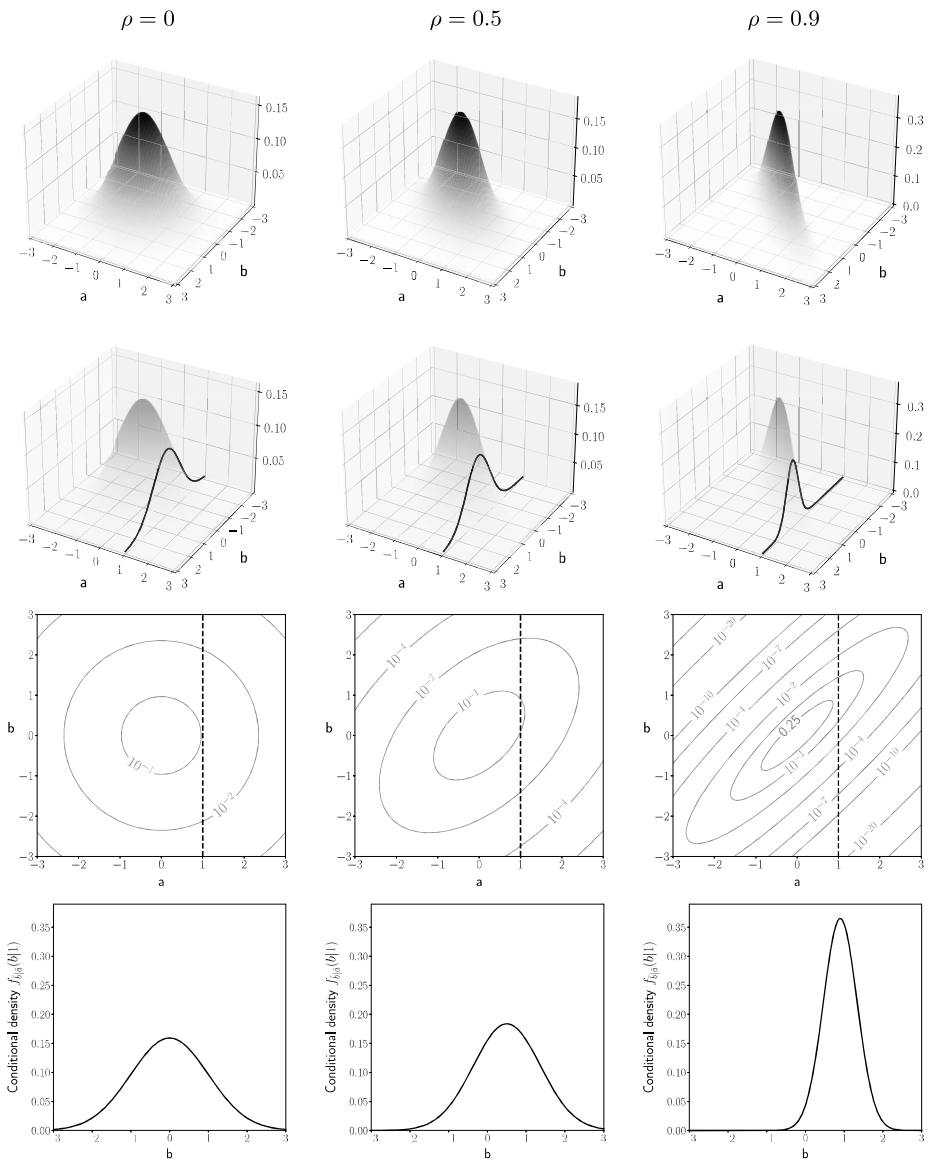

Figure 5.17 Conditional distributions of Gaussian random vectors. The first row shows the joint pdf of the Gaussian random vector in Example 5.23 for different values of the correlation coefficient ρ. In the second row we see that the slice of the density corresponding to $a = 1$ has a Gaussian shape. The third row shows the contour plots and a dashed line at $a = 1$. The fourth row shows the corresponding conditional pdf of $\tilde{b}$, which is indeed Gaussian with mean ρa and variance $1 - \rho^2$.

In Example 5.23, the marginal and conditional pdf of the Gaussian random vector are also Gaussian. This is no accident. The following theorem shows that the marginal and conditional distributions of the entries of any 2D Gaussian random vector are Gaussian.

Theorem 5.24 (Marginal and conditional distributions of a 2D Gaussian)**.** *Let $(\tilde{a}, \tilde{b})$ be a Gaussian random vector with mean $\mu \in \mathbb{R}^2$ and covariance matrix $\Sigma \in \mathbb{R}^{2\times 2}$, where Σ is positive definite and full rank. We reparametrize the mean and covariance, as follows*

$$\mu := \begin{bmatrix} \mu_{\tilde{a}} \\ \mu_{\tilde{b}} \end{bmatrix}, \tag{5.119}$$

$$\Sigma := \begin{bmatrix} \sigma_{\tilde{a}}^2 & \rho\sigma_{\tilde{a}}\sigma_{\tilde{b}} \\ \rho\sigma_{\tilde{a}}\sigma_{\tilde{b}} & \sigma_{\tilde{b}}^2 \end{bmatrix}, \tag{5.120}$$

where $-1 < \rho < 1$ (otherwise Σ is not positive definite, which can be verified using the eigendecomposition of Σ).

The marginal distribution of $\tilde{a}$ is Gaussian with mean $\mu_{\tilde{a}}$ and standard deviation $\sigma_{\tilde{a}}$. The marginal distribution of $\tilde{b}$ is Gaussian with mean $\mu_{\tilde{b}}$ and standard deviation $\sigma_{\tilde{b}}$. The conditional distribution of $\tilde{b}$ given $\tilde{a} = a$ is Gaussian with mean

$$\mu_{\text{cond}} := \mu_{\tilde{b}} + \frac{\rho\sigma_{\tilde{b}}(a - \mu_{\tilde{a}})}{\sigma_{\tilde{a}}} \tag{5.121}$$

and variance

$$\sigma_{\text{cond}}^2 := (1 - \rho^2)\sigma_{\tilde{b}}^2. \tag{5.122}$$

Proof The proof follows the exact same reasoning as in Example 5.23. The inverse of the covariance matrix is

$$\Sigma^{-1} = \frac{1}{\sigma_{\tilde{a}}^2\sigma_{\tilde{b}}^2(1-\rho^2)} \begin{bmatrix} \sigma_{\tilde{b}}^2 & -\rho\sigma_{\tilde{a}}\sigma_{\tilde{b}} \\ -\rho\sigma_{\tilde{a}}\sigma_{\tilde{b}} & \sigma_{\tilde{a}}^2 \end{bmatrix}. \tag{5.123}$$

We define

$$s(a) := \frac{a - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}}, \tag{5.124}$$

$$s(b) := \frac{b - \mu_{\tilde{b}}}{\sigma_{\tilde{b}}}. \tag{5.125}$$

This is often called *standardizing* the variables (see Section 8.2). We then have

$$f_{\tilde{a},\tilde{b}}(a,b) = \frac{1}{2\pi\sigma_{\tilde{a}}\sigma_{\tilde{b}}\sqrt{1-\rho^2}} \exp\left(-\frac{s(a)^2 - 2\rho s(a)s(b) + s(b)^2}{2(1-\rho^2)}\right) \tag{5.126}$$

$$= \frac{1}{2\pi\sigma_{\tilde{a}}\sigma_{\tilde{b}}\sqrt{1-\rho^2}} \exp\left(-\frac{(1-\rho^2)s(a)^2 + (s(b) - \rho s(a))^2}{2(1-\rho^2)}\right) \tag{5.127}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_{\tilde{a}}} \exp\left(-\frac{s(a)^2}{2}\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_{\tilde{b}}} \exp\left(-\frac{(s(b) - \rho s(a))^2}{2(1-\rho^2)}\right)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_{\tilde{a}}} \exp\left(-\frac{(a - \mu_{\tilde{a}})^2}{2\sigma_{\tilde{a}}^2}\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_{\tilde{b}}} \exp\left(-\frac{(b - \mu_{\tilde{b}} - \rho\sigma_{\tilde{b}}s(a))^2}{2(1-\rho^2)\sigma_{\tilde{b}}^2}\right).$$

As in Example 5.23, we realize that the second term is Gaussian with respect to b, so it must integrate to one. This means that the other term must be the marginal pdf of $\tilde{a}$. Dividing the joint pdf by the marginal pdf then establishes that the second term is the conditional pdf of $\tilde{b}$. The marginal pdf of $\tilde{b}$ can be derived by switching the roles of a and b. ■

Theorem 5.24 shows that each component of a Gaussian 2D vector is Gaussian, and so are their conditional distributions. The correlation coefficient ρ again governs the dependence between the two components. We elaborate further on the properties of the correlation coefficient in Chapter 8.

Our results for 2D Gaussian vectors generalize to higher dimensions. The marginal and conditional distributions of any subvector of a Gaussian random vector are also Gaussian.

Theorem 5.25 (Marginal and conditional distributions of multidimensional Gaussian)**.** *Let $\tilde{z}$ denote a d-dimensional Gaussian random vector with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, which is positive definite and full rank. Assume that the subvectors $\tilde{x}$ and $\tilde{y}$ consist of the first m and $d-m$ entries respectively for any $1 \leq m < d$,*

$$\tilde{z} := \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix}. \tag{5.128}$$

This is without loss of generality; we can reorder the entries of any random vector to place a subvector of interest at the beginning. We express μ and Σ as

$$\mu := \begin{bmatrix} \mu_{\tilde{x}} \\ \mu_{\tilde{y}} \end{bmatrix}, \tag{5.129}$$

$$\Sigma_{\tilde{z}} := \begin{bmatrix} \Sigma_{\tilde{x}} & \Sigma_{\tilde{x},\tilde{y}} \\ \Sigma_{\tilde{x},\tilde{y}}^T & \Sigma_{\tilde{y}} \end{bmatrix}, \tag{5.130}$$

where $\mu_{\tilde{x}} \in \mathbb{R}^m$, $\mu_{\tilde{y}} \in \mathbb{R}^{d-m}$, $\Sigma_{\tilde{x}} \in \mathbb{R}^{m \times m}$, $\Sigma_{\tilde{x},\tilde{y}} \in \mathbb{R}^{m \times (d-m)}$, $\Sigma_{\tilde{y}} \in \mathbb{R}^{(d-m) \times (d-m)}$.

The marginal distribution of $\tilde{x}$ is Gaussian with mean $\mu_{\tilde{x}}$ and covariance matrix $\Sigma_{\tilde{x}}$. The marginal distribution of $\tilde{y}$ is Gaussian with mean $\mu_{\tilde{y}}$ and covariance matrix $\Sigma_{\tilde{y}}$. The conditional distribution of $\tilde{y}$ given $\tilde{x} = x$ is Gaussian with mean

$$\mu_{\text{cond}} = \mu_{\tilde{y}} + \Sigma_{\tilde{x},\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} (x - \mu_{\tilde{x}}) \tag{5.131}$$

and covariance matrix

$$\Sigma_{\text{cond}} = \Sigma_{\tilde{y}} - \Sigma_{\tilde{x},\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x},\tilde{y}}. \tag{5.132}$$

Proof The proof follows the same argument as the proof of Theorem 5.24. We factorize the joint pdf of $\tilde{z}$ into two terms,

$$\begin{aligned} f_{\tilde{z}} \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) = & \frac{1}{\sqrt{(2\pi)^m \left| \Sigma_{\tilde{x}} \right|}} \exp \left(-\frac{1}{2} \left(x - \mu_{\tilde{x}} \right)^T \Sigma_{\tilde{x}}^{-1} \left(x - \mu_{\tilde{x}} \right) \right) \\ & \frac{1}{\sqrt{(2\pi)^{d-m} \left| \Sigma_{\text{cond}} \right|}} \exp \left(-\frac{1}{2} \left(y - \mu_{\text{cond}} \right)^T \Sigma_{\text{cond}}^{-1} \left(y - \mu_{\text{cond}} \right) \right). \end{aligned} \tag{5.133}$$

This can be achieved through algebraic manipulations that rely on the Schur complement, which we omit as they are lengthy and not particularly insightful. We refer the reader to (**?**) for a detailed derivation. The second term in (5.133) is the joint pdf of a Gaussian random vector, which integrates to one. The first term is therefore the marginal pdf of $\tilde{x}$. The marginal pdf of $\tilde{y}$ can be derived by switching the roles of $\tilde{x}$ and $\tilde{y}$. Dividing the joint pdf by the first term establishes that the second term is the conditional pdf of $\tilde{y}$ given $\tilde{x} = x$. ■

5.10.3 Maximum-Likelihood Estimation

In order to fit a multidimensional Gaussian distribution to a d-dimensional dataset $X := \{x_1, \ldots, x_n\}$, we apply the maximum-likelihood method described in Section 3.7. This requires maximizing the likelihood of the data with respect to the mean and covariance parameters. Assuming i.i.d. data sampled from a Gaussian distribution with unknown parameters μ and Σ, the likelihood is

$$\mathcal{L}(\mu, \Sigma) := \prod_{i=1}^{n} f_{\mu,\Sigma}(x_i) \tag{5.134}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d \, |\Sigma|}} \exp\left(-\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right), \tag{5.135}$$

and the log-likelihood is

$$\log \mathcal{L}(\mu, \Sigma) := -n \log \sqrt{(2\pi)^d \, |\Sigma|} - \sum_{i=1}^{n} \frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu). \tag{5.136}$$

The following theorem provides the maximum-likelihood of the parameters of the Gaussian random vector. The mean parameter is obtained by averaging the data, and is therefore equal to the sample mean (see Definition 11.4). The covariance matrix is obtained by averaging the outer product of each data point with itself, after subtracting the mean, which is essentially equal to the sample covariance matrix (see Definition 11.14).∗

Theorem 5.26 (Maximum-likelihood estimation for Gaussian random vectors)**.** *Let $X := \{x_1, \ldots, x_n\}$ be n data in $\mathbb{R}^d$. The maximum-likehood estimates of the parameters of a Gaussian random vector given these data are*

$$\mu_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{5.137}$$

$$\Sigma_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{\text{ML}})(x_i - \mu_{\text{ML}})^T. \tag{5.138}$$

∗There is a minor discrepancy in the scaling factor, which is set to $n - 1$ in Definition 11.14 so that the estimator is unbiased, but unless n is very small this doesn't really make any difference in practice.

Proof For fixed Σ, the gradient and Hessian of the log likelihood with respect to μ equal

$$\nabla_\mu \log \mathcal{L}(\mu, \Sigma) = \Sigma^{-1} \sum_{i=1}^{n} (x_i - \mu), \tag{5.139}$$

$$\nabla^2_\mu \log \mathcal{L}(\mu, \Sigma) = -\Sigma^{-1}. \tag{5.140}$$

Since Σ is positive definite, Σ^{-1} is also positive definite (the eigenvalues of Σ^{-1} are just the inverses of the eigenvalues of Σ). This means that the log likelihood is concave with respect to μ and can be maximized by setting the gradient to zero, which yields (5.137). Maximizing with respect to the covariance matrix is significantly more complicated. We plug in μ_{ML}, since it doesn't depend on the value of Σ, and take the gradient of log likelihood with respect to Σ^{-1},

$$\nabla_{\Sigma^{-1}} \log \mathcal{L}(\mu, \Sigma) = \frac{n}{2}\Sigma - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T. \tag{5.141}$$

Deriving the gradient requires some pretty lengthy matrix computations, which we omit. We refer the reader to (**?**) for the gory details. ■

In order to illustrate the application of the Gaussian parametric model to real data, we use it to model the height and weight, and the height and foot length, of a group of US army members, extracted from Dataset 5. The joints pdfs, obtained by computing the maximum-likelihood parameter estimates in Theorem 5.26, are shown in Figure 5.18. They are very similar to the density estimates produced by kernel density estimation, indicating that the distributions are indeed approximately Gaussian.

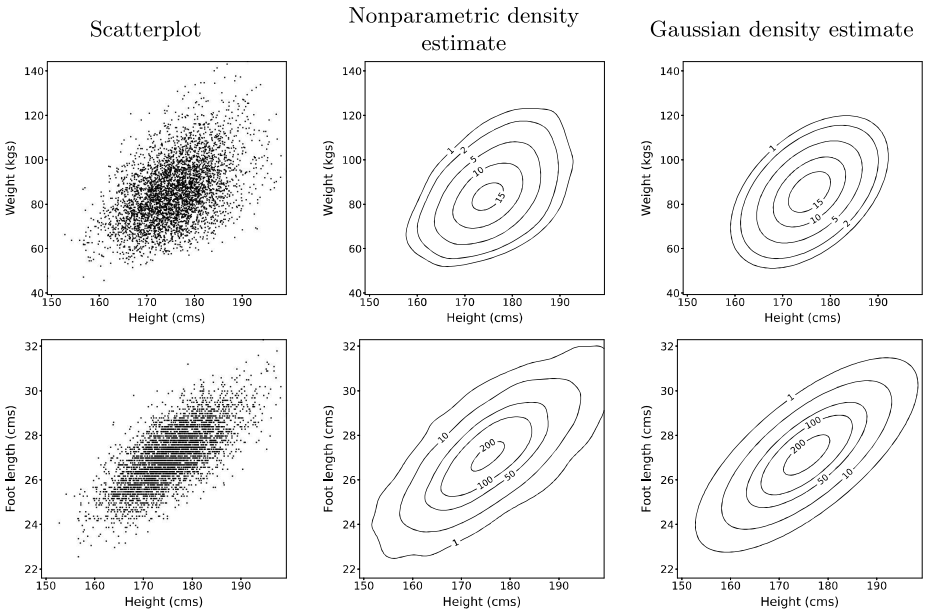

Figure 5.18 Gaussian parametric modeling of height, weight and foot length. The first column shows the scatterplot of the height and weight (first row) and the height and foot length (second row) of 4,082 men in the United States army, extracted from Dataset 5. The second column shows the contour lines of the corresponding nonparametric estimate of the joint density obtained via KDE. The third column shows the density of a parametric Gaussian model fit via maximum likelihood.