

# Chapter 5

## Modeling Dependence

To study relationships between random variables we need to add more concepts to our toolkit. In this chapter we review multivariate distributions and give a rigorous construction of conditional expectation.

### 5.1 Random Vectors and Matrices

Let's begin with random vectors and matrices.

#### 5.1.1 Random Vectors

In §4.1.3 we defined a random variable to be a function  $x: \Omega \rightarrow \mathbb{R}$  with the property that  $\{x \in B\}$  is an event for any Borel set  $B \in \mathcal{B}(\mathbb{R})$ . The definition of a random vector is analogous. A **random vector**  $\mathbf{x}$  in  $\mathbb{R}^N$  is a function from  $\Omega$  to  $\mathbb{R}^N$  with the property that

$$\{\omega \in \Omega : \mathbf{x}(\omega) \in B\} \in \mathcal{F} \quad \text{for all } B \in \mathcal{B}(\mathbb{R}^N) \quad (5.1)$$

This is just (4.9) on page 92 with  $\mathcal{B}(\mathbb{R}^N)$  replacing  $\mathcal{B}(\mathbb{R})$ . In statistical applications we can almost always cook up a sample space  $\Omega$  and a set of events  $\mathcal{F}$  on  $\Omega$  such that all random vectors we care about satisfy (5.1).<sup>1</sup>

It's also possible to define a **random vector**  $\mathbf{x}$  in  $\mathbb{R}^N$  as a tuple of  $N$  random variables  $(x_1, \dots, x_N)$ . The two definitions are equivalent. As with ordinary vectors, random vectors have no intrinsic shape, and we'll write them in rows or columns according to convenience. However, in expressions involving matrix multiplication, random vectors will default to column vectors unless otherwise stated.

---

1. For technical details on these constructions see, for example, chapter 4 of Çinlar (2011).

**Example 5.1.1** Recall the blindfolded monkey experiment from page 80. The sample space is the unit disk  $\Omega := \{(h, v) \in \mathbb{R}^2 : \|(h, v)\| \leq 1\}$  and the event space is the Borel sets in  $\Omega$ . If  $\mathbf{x}$  is the identity on  $\Omega$ , then it simply reports the outcome  $(h, v)$ . This is one instance of a random vector.

**Example 5.1.2** Consider a random sample listing the income  $y_n$  of  $n = 1, \dots, N$  individuals from a given population. The vector  $(y_1, \dots, y_N)$  that reports the outcome of this sampling can be regarded as a random vector in  $\mathbb{R}^N$ .

The definition of a random vector ensures that  $\{\mathbf{x} \in B\}$  is a well-defined event for every  $B \in \mathcal{B}(\mathbb{R}^N)$ . If we want to ensure that some transformation  $\mathbf{y} = f(\mathbf{x})$  is likewise a random vector, we need a restriction on  $f$  so that  $\{\mathbf{y} \in B\}$  is also a well-defined event. The appropriate restriction is  $\mathcal{B}$ -measurability, which, analogous to the scalar case (see page 92), means that  $f: \mathbb{R}^N \rightarrow \mathbb{R}^M$  satisfies  $f^{-1}(B) \in \mathcal{B}(\mathbb{R}^N)$  for all  $B \in \mathcal{B}(\mathbb{R}^M)$ . As in the scalar case, every kind of transformation we use in this text is  $\mathcal{B}$ -measurable, so the restriction is never binding.

Expectations of random vectors are defined element-by-element. In particular, if  $\mathbf{x} = (x_1, \dots, x_N)$  is a random vector in  $\mathbb{R}^N$ , then its **expectation**  $\mathbb{E}\mathbf{x}$  is defined as the vector of expectations

$$\mathbb{E}\mathbf{x} = \mathbb{E} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} := \begin{pmatrix} \mathbb{E}x_1 \\ \mathbb{E}x_2 \\ \vdots \\ \mathbb{E}x_N \end{pmatrix}$$

An  $M \times N$  **random matrix**  $\mathbf{X}$  is an  $M \times N$  array of random variables. Its expectation is defined as

$$\mathbb{E}\mathbf{X} := \begin{pmatrix} \mathbb{E}x_{11} & \cdots & \mathbb{E}x_{1N} \\ \vdots & & \vdots \\ \mathbb{E}x_{M1} & \cdots & \mathbb{E}x_{MN} \end{pmatrix}$$

**Fact 5.1.1** If  $\mathbf{X}$  and  $\mathbf{Y}$  are random matrices or vectors and  $\mathbf{A}$  and  $\mathbf{B}$  are constant and conformable, then

$$\mathbb{E}[\mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{Y}] = \mathbf{A}\mathbb{E}[\mathbf{X}] + \mathbf{B}\mathbb{E}[\mathbf{Y}]$$

Fact 5.1.1 is obtained by repeated application of the linearity of scalar expectations stated in fact 4.1.5.

The **variance–covariance matrix** of a random vector  $\mathbf{x}$  in  $\mathbb{R}^N$  with  $\boldsymbol{\mu} := \mathbb{E}\mathbf{x}$  is the  $N \times N$  matrix

$$\text{var}[\mathbf{x}] := \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] \quad (5.2)$$

Expanding this out gives

$$\text{var}[\mathbf{x}] = \begin{pmatrix} \mathbb{E}[(x_1 - \mu_1)(x_1 - \mu_1)] & \cdots & \mathbb{E}[(x_1 - \mu_1)(x_N - \mu_N)] \\ \vdots & & \vdots \\ \mathbb{E}[(x_N - \mu_N)(x_1 - \mu_1)] & \cdots & \mathbb{E}[(x_N - \mu_N)(x_N - \mu_N)] \end{pmatrix}$$

The  $j, k$ th term is the scalar covariance between  $x_j$  and  $x_k$  and the principal diagonal contains the variance of each  $x_n$ .

**Fact 5.1.2** For any random vector  $\mathbf{x}$  with  $\mathbb{E}[\mathbf{x}^\top \mathbf{x}] < \infty$ ,

- (i)  $\text{var}[\mathbf{x}]$  exists and is nonnegative definite,
- (ii)  $\text{var}[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mu\mu^\top$ , and
- (iii)  $\text{var}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A} \text{var}[\mathbf{x}] \mathbf{A}^\top$  (any  $\mathbf{A}, \mathbf{b}$  constant and conformable).

The **cross-covariance** between random vectors  $\mathbf{x}$  and  $\mathbf{y}$  is defined as

$$\text{cov}[\mathbf{x}, \mathbf{y}] := \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^\top]$$

Evidently  $\text{var}[\mathbf{x}] = \text{cov}[\mathbf{x}, \mathbf{x}]$ .

Here's a useful fact concerning expectations of quadratic forms. The proof is a solved exercise (see ex. 5.4.7).

**Fact 5.1.3** If  $\mathbf{z}$  is a random vector in  $\mathbb{R}^N$  satisfying  $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \mathbf{I}$  and  $\mathbf{A}$  is any constant  $N \times N$  matrix, then

$$\mathbb{E}[\mathbf{z}^\top \mathbf{A} \mathbf{z}] = \text{trace } \mathbf{A}$$

## 5.1.2 Multivariate Distributions

The definition of a distribution over  $\mathbb{R}^N$  generalizes that of a distribution over  $\mathbb{R}$  given in 4.2.1. In particular, a **distribution** or **law**  $P$  on  $\mathbb{R}^N$  is a probability measure over the Borel sets  $\mathcal{B}(\mathbb{R}^N)$ . By definition, it satisfies  $P(\mathbb{R}^N) = 1$  and  $P(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} P(B_n)$  for any disjoint sequence  $\{B_n\}$  in  $\mathcal{B}(\mathbb{R}^N)$ . See also the properties of probability measures listed in fact 4.1.1 on page 86.

Figure 5.1 shows a visualization of a distribution  $P$  on  $\mathbb{R}^2$ , with hot colors indicating regions of high probability. (The distribution is multivariate Gaussian, to be defined below.) The sets  $A$  and  $B$  are typical Borel sets. In the figure,  $B$  has the same area as  $A$  but is assigned greater probability by  $P$ . By additivity, we know that  $P(A \cup B) = P(A) + P(B)$  must hold.

As was the case for  $\mathbb{R}$ , distributions on  $\mathbb{R}^N$  are fully characterized by their values on certain more elementary families of subsets of  $\mathbb{R}^N$ . For example, for any two

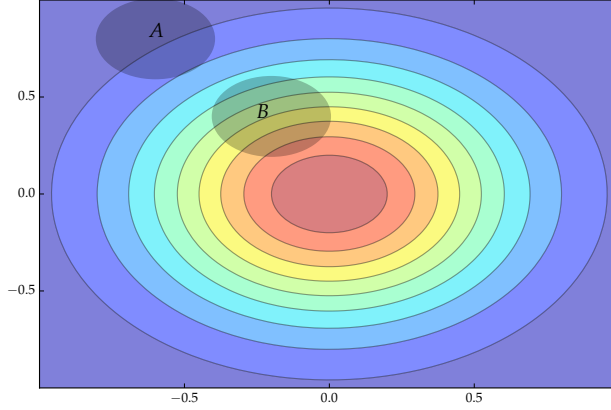


**Figure 5.1** Example distribution and events  $A$  and  $B$

distributions  $Q, P$  on  $\mathbb{R}^N$ ,

$$P(I) = Q(I) \text{ for all } I \in \mathcal{R} \implies P(B) = Q(B) \text{ for all } B \in \mathcal{B}(\mathbb{R}^N)$$

where  $\mathcal{R}$  denotes the family of all **cylinder sets**

$$C = B_1 \times \cdots \times B_N, \quad B_n \in \mathcal{B}(\mathbb{R}), \quad n = 1, \dots, N$$

or, specializing further, the family of all rectangles<sup>2</sup>

$$I = (-\infty, s_1] \times \cdots \times (-\infty, s_N], \quad s_n \in \mathbb{R}, \quad n = 1, \dots, N$$

(See prop 3.7 of Çinlar 2011.) Figure 5.2 shows an example set  $I = (-\infty, s_1] \times (-\infty, s_2]$  in the case of  $\mathbb{R}^2$ .

Thus, analogous to (4.24), any distribution  $P$  on  $\mathbb{R}^N$  is characterized by the function

$$F(\mathbf{s}) := F(s_1, \dots, s_N) := P\left(\times_{n=1}^N (-\infty, s_n]\right) \quad (\mathbf{s} \in \mathbb{R}^N) \quad (5.3)$$

This is an example of a **multivariate cumulative distribution function**, which is a function  $F: \mathbb{R}^N \rightarrow [0, 1]$  that is

- (i) right-continuous in each of its arguments,

---

2. See page 408 for more discussion of rectangles.

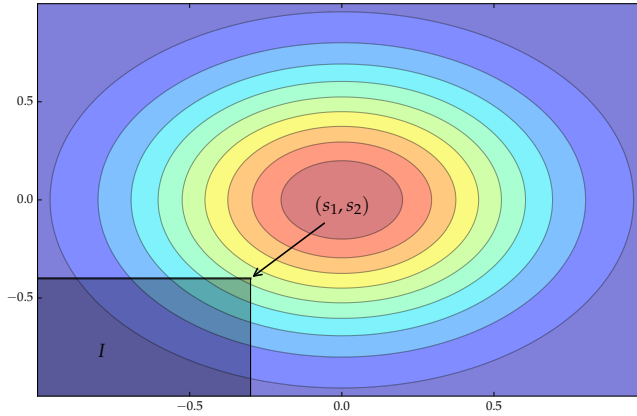


**Figure 5.2** Distributions are characterized by their values on elements of  $\mathcal{B}$

- (ii) increasing in each of its arguments, and
- (iii) satisfies

$$F(\mathbf{s}_j) \rightarrow 1 \text{ as } \mathbf{s}_j \rightarrow \infty \quad \text{and} \quad F(s_1, \dots, s_{nj}, \dots, s_N) \rightarrow 0 \text{ as } s_{nj} \rightarrow -\infty$$

As in the one-dimensional case, we will call a distribution  $P$  on  $\mathbb{R}^N$  **discrete** if  $P$  is supported on a countable subset of  $\mathbb{R}^N$ , and **absolutely continuous** if  $P(B) = 0$  whenever  $B$  has zero Lebesgue measure. The latter condition is again necessary and sufficient for existence of a density representation, by which we mean existence of a density  $p$  on  $\mathbb{R}^N$  such that

$$P(B) = \int_B p(\mathbf{s}) \, d\mathbf{s} \quad \text{for all } B \in \mathcal{B}(\mathbb{R}^N) \quad (5.4)$$

The right-hand side of (5.4) is a multivariate integral that we can write out more explicitly as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{1}_B(s_1, \dots, s_N) p(s_1, \dots, s_N) \, ds_1 \cdots ds_N \quad (5.5)$$

Conversely, if  $p$  is any density on  $\mathbb{R}^N$ , then (5.4) defines a distribution.

**Example 5.1.3** The **multivariate normal density** or **multivariate Gaussian density** on

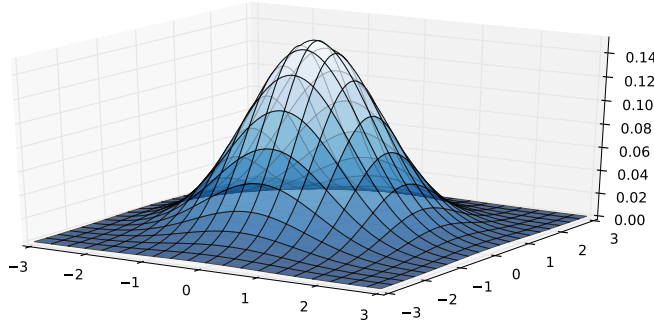


**Figure 5.3** Bivariate standard normal density

$\mathbb{R}^N$  is a function  $p$  of the form

$$p(\mathbf{s}) = (2\pi)^{-N/2} \det(\mathbf{\Sigma})^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{s} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{s} - \boldsymbol{\mu}) \right\} \quad (5.6)$$

where  $\boldsymbol{\mu}$  is any  $N \times 1$  vector and  $\mathbf{\Sigma}$  is a positive definite  $N \times N$  matrix. In symbols, we represent this distribution by  $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ . The case  $N(\mathbf{0}, \mathbf{I})$  is called the **multivariate standard normal distribution**. Figure 5.3 shows the standard normal density when  $N = 2$ .

Let's describe one important way to construct multivariate distributions.

**Fact 5.1.4** Given distributions  $P_1, \dots, P_N$  on  $\mathbb{R}$ , there exists a unique and well-defined distribution  $\mathring{P}$  on  $\mathbb{R}^N$  such that

$$\mathring{P}(B_1 \times \cdots \times B_N) = \prod_{n=1}^N P_n(B_n) \quad \text{for all } B_n \in \mathcal{B}(\mathbb{R}), n = 1, \dots, N$$

This distribution is called the **product distribution** of  $P_1, \dots, P_N$ . If  $P_n = P$  for all  $n$ , then we'll call  $\mathring{P}$  the  **$N$ th product of  $P$** . The uniqueness part of fact 5.1.4 follows from the claim above that distributions on  $\mathbb{R}^N$  are uniquely pinned down by their values on cylinder subsets of  $\mathbb{R}^N$ . As we'll see, product distributions correspond to joint distributions of independent random variables.

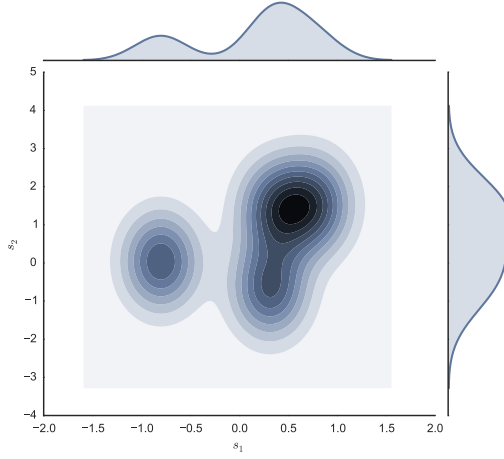


**Figure 5.4** Bivariate joint density and its two marginals

#### 5.1.2.1 Marginals of Multivariate Distributions

Given any distribution  $P$  on  $\mathbb{R}^N$ , the  $n$ th **marginal distribution** of  $P$  is the distribution on  $\mathbb{R}$  defined by

$$P_n(B) = P(\mathbb{R} \times \cdots \times \mathbb{R} \times B \times \mathbb{R} \times \cdots \times \mathbb{R})$$

Here  $B$  is the  $n$ th element of the Cartesian product. We can also write this as

$$P_n(B) = P\{\mathbf{s} \in \mathbb{R}^N : \mathbf{s}^\top \mathbf{e}_n \in B\}$$

The  $n$ th marginal tells us the probabilistic implications of  $P$  when we only concern ourselves with the outcomes for the  $n$ th coordinate. From  $P_n$  we can also extract the **marginal CDF**  $F_n$  via (4.24) on page 100.

If  $P_n$  is absolutely continuous, it has density  $p_n$ . When the joint distribution  $P$  has a density  $p$ , the marginal distribution  $P_n$  has a density, and  $p_n$  can be obtained directly from  $p$  by integrating out the other variables. For example, in the bivariate case,

$$p_1(s_1) = \int_{-\infty}^{\infty} p(s_1, s_2) ds_2$$

Figure 5.4 shows a bivariate density represented by contours and the two marginals obtained from projecting onto each axis.

In general, the joint distribution cannot be determined from the marginals alone, since the marginals do not tell us about the interactions across coordinates. The exception is when there is no interaction. This is the case for product distributions, as described in fact 5.1.4 and further discussed in §5.1.4.

### 5.1.3 Distributions of Random Vectors

Let  $\mathbf{x}$  be a random vector in  $\mathbb{R}^N$ . The **distribution** of  $\mathbf{x}$  is the probability measure  $P$  on  $\mathcal{B}(\mathbb{R}^N)$  defined by

$$P(B) = \mathbb{P}\{\mathbf{x} \in B\} \quad (B \in \mathcal{B}(\mathbb{R}^N)) \quad (5.7)$$

In this context,  $P$  is also called the **joint distribution** of  $x_1, \dots, x_N$ , and we write  $\mathcal{L}(\mathbf{x}) = P$ . The joint distribution can also be represented by the multivariate CDF  $F: \mathbb{R}^N \rightarrow [0, 1]$  defined by (5.3), which in this case can be written as  $F(s_1, \dots, s_N) = \mathbb{P}\{x_1 \leq s_1, \dots, x_N \leq s_N\}$ , or, in vector notation,

$$F(\mathbf{s}) = \mathbb{P}\{\mathbf{x} \leq \mathbf{s}\} \quad (\mathbf{s} \in \mathbb{R}^N)$$

It follows from (5.4) that when the distribution  $P$  of  $\mathbf{x}$  is absolutely continuous, there exists a nonnegative function  $p$  on  $\mathbb{R}^N$  satisfying

$$\int_B p(\mathbf{s}) \, d\mathbf{s} = \mathbb{P}\{\mathbf{x} \in B\} \quad (B \in \mathcal{B}(\mathbb{R}^N)) \quad (5.8)$$

Here  $\int_B$  is a multivariate integral over the region  $B$ . The function  $p$  is called either the density or **joint density** of  $\mathbf{x}$ . For (5.8) to hold, it suffices that

$$\int_{-\infty}^{s_N} \cdots \int_{-\infty}^{s_1} p(t_1, \dots, t_N) \, dt_1 \cdots dt_N = F(s_1, \dots, s_N) \quad (5.9)$$

for all  $s_n \in \mathbb{R}$ ,  $n = 1, \dots, N$ .

#### 5.1.3.1 Joint versus Marginal

If  $\mathbf{x} = (x_1, \dots, x_N)$  is a random vector in  $\mathbb{R}^N$ , then each  $x_n$  is a random variable on  $\mathbb{R}$ . Let  $P_n = \mathcal{L}(x_n)$ , so that

$$P_n(B) = \mathbb{P}\{x_n \in B\} \quad (B \in \mathcal{B}(\mathbb{R}), n = 1, \dots, N) \quad (5.10)$$

This is of course the marginal distribution relative to the joint distribution of  $\mathbf{x}$ , as introduced in §5.1.2.1. In this context it is called the **marginal distribution of  $x_n$**  in order to distinguish it from the (joint) distribution of  $\mathbf{x}$ . From a mathematical perspective, the marginal distribution of  $x_n$  and the distribution of  $x_n$  are identical.

If  $P_1 = P_2 = \cdots = P_N$ , then we say that  $x_1, \dots, x_N$  are **identically distributed**.

### 5.1.3.2 Gaussian Random Vectors

Let  $z$  be standard normal on  $\mathbb{R}$ . A random variable  $x$  is said to be **normally distributed** if  $x = \mu + \sigma z$  for some  $\sigma \geq 0$ , and we write  $\mathcal{L}(x) = N(\mu, \sigma)$ . Here we explicitly allow the possibility that  $\sigma = 0$ , in which case  $\text{var}[x] = 0$  and  $\mathbb{P}\{x = \mu\} = 1$ .

A random vector  $\mathbf{x}$  in  $\mathbb{R}^N$  is called **multivariate normal** or **multivariate Gaussian** if there exists a  $K \times 1$  standard normal random vector  $\mathbf{z}$ , an  $N \times K$  matrix  $\mathbf{C}$  and an  $N \times 1$  vector  $\boldsymbol{\mu}$  such that

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{C}\mathbf{z} \quad (5.11)$$

Once again we explicitly allow degenerate cases, since they occur frequently in statistics. If  $\mathbf{x}$  is multivariate normal, then we write  $\mathcal{L}(\mathbf{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , where

$$\boldsymbol{\mu} := \mathbb{E}\mathbf{x} \quad \text{and} \quad \boldsymbol{\Sigma} := \text{var } \mathbf{x}$$

In the case of (5.11) we have  $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\top$ . See fact 5.1.2 on page 127.

Be aware that  $\mathcal{L}(\mathbf{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  does not imply that  $\mathbf{x}$  has the multivariate normal density in (5.6). The reason is that the distribution of  $\mathbf{x}$  can fail to be absolutely continuous. For example, if  $\mathbf{C} = \mathbf{0}$  in (5.11), then  $\mathbf{x}$  puts all mass on a single point in  $\mathbb{R}^N$ . Without absolute continuity, no density representation can exist.

Absolute continuity of the distribution of  $\mathbf{x}$  coincides with the setting where  $\boldsymbol{\Sigma} := \text{var } \mathbf{x}$  is nonsingular. (Perhaps because of this, the scenario where a random vector has a density is sometimes called the “nonsingular case” by econometricians.) For  $\mathbf{x}$  defined in (5.11), nonsingularity of  $\boldsymbol{\Sigma}$  will be true if and only if  $\mathbf{C}^\top$  has full column rank (see ex. 3.5.28 on page 72).

**Fact 5.1.5** Let  $\mathbf{x}$  be a random vector in  $\mathbb{R}^N$ . The following statements are true:

- (i) The vector  $\mathbf{x}$  is multivariate normal if and only if  $\mathbf{a}^\top \mathbf{x}$  is normally distributed in  $\mathbb{R}$  for every constant  $N \times 1$  vector  $\mathbf{a}$ .
- (ii) If  $\mathcal{L}(\mathbf{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , then

$$\mathcal{L}(\mathbf{A}\mathbf{x} + \mathbf{b}) = N(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$$

for all constant conformable  $\mathbf{A}, \mathbf{b}$ .

One corollary of fact 5.1.5 is that if  $\mathbf{x} = (x_1, \dots, x_N)$  is multivariate normal, then the marginal distribution of  $x_n$  is univariate normal. This can be seen by taking  $\mathbf{a} = \mathbf{e}_n$  in part (i). In fact the rule in part (ii) implies that

$$\mathcal{L}(x_n) = \mathcal{L}(\mathbf{e}_n^\top \mathbf{x}) = N(\mathbf{e}_n^\top \boldsymbol{\mu}, \mathbf{e}_n^\top \boldsymbol{\Sigma} \mathbf{e}_n) = N(\mu_n, \sigma_n^2) \quad (5.12)$$

where  $\sigma_n^2$  is the  $n, n$ th element of  $\Sigma$ .

Here's a common exam question: Is it true that the joint distribution of  $N$  univariate normal random variables is always multivariate normal? The answer is no. A counterexample is given in §5.1.5. See also fact 5.1.14 below.

### 5.1.3.3 Expectations from Distributions

Let  $h: \mathbb{R}^N \rightarrow \mathbb{R}$  be any  $\mathcal{B}$ -measurable function and let  $P$  be a distribution on  $\mathbb{R}^N$ . In §4.2.3 we saw that one can assign meaning to the expression  $\int h(s)P(ds)$  by treating  $h$  as a random variable on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$  and taking its expectation. We can do exactly the same thing in  $\mathbb{R}^N$ , with  $h$  now regarded as a random variable on  $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N), P)$ . Its expectation will be written as

$$\mathbb{E}_P h := \int h(\mathbf{s})P(d\mathbf{s}) \quad (5.13)$$

Here's a multivariate version of fact 4.2.3 on page 110:

**Fact 5.1.6** Let  $h: \mathbb{R}^N \rightarrow \mathbb{R}$  be  $\mathcal{B}$ -measurable and let  $P$  be a distribution on  $\mathbb{R}^N$ . If  $P$  is discrete, with PMF  $\{p_j\}_{j \geq 1}$  and support  $\{\mathbf{s}_j\}_{j \geq 1}$ , then

$$\int h(\mathbf{s})P(d\mathbf{s}) = \sum_{j \geq 1} h(\mathbf{s}_j)p_j \quad (5.14)$$

If  $P$  is absolutely continuous with density  $p$ , then

$$\int h(\mathbf{s})P(d\mathbf{s}) = \int h(\mathbf{s})p(\mathbf{s}) d\mathbf{s} \quad (5.15)$$

The right-hand side of (5.15) should be understood as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(s_1, \dots, s_N) p(s_1, \dots, s_N) ds_1 \cdots ds_N$$

In §4.2.5 we saw that expectations of functions of random variables can be computed from their distributions. The same is true for functions of random vectors. In particular, suppose that we want to compute  $\mathbb{E}h(\mathbf{x})$ , where  $h: \mathbb{R}^N \rightarrow \mathbb{R}$  is  $\mathcal{B}$ -measurable. We can do this either by evaluating the expectation of  $h(\mathbf{x})$  on the original probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , or we can take the distribution  $P$  of  $\mathbf{x}$  and compute the expectation of  $h$  on  $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N), P)$ . This is formalized in the following extension to fact 4.2.8 on page 112.

**Fact 5.1.7** If  $\mathcal{L}(\mathbf{x}) = P$  and  $h: \mathbb{R}^N \rightarrow \mathbb{R}$  is a  $\mathcal{B}$ -measurable function such that  $h(\mathbf{x})$  is integrable, then

$$\mathbb{E}h(\mathbf{x}) = \int h(\mathbf{s})P(d\mathbf{s})$$

where  $\int h(\mathbf{s})P(d\mathbf{s})$  is the expectation of  $h$  on  $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N), P)$ .

As in the univariate case, this means that objects like moments are properties of the distribution. For example, let  $\mathbf{x}$  be a random vector in  $\mathbb{R}^K$  with  $\mathcal{L}(\mathbf{x}) = P$ . The variance–covariance matrix  $\text{var}[\mathbf{x}]$  of  $\mathbf{x}$  has  $i, j$ th element  $\mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j]$ . We can write this in terms of  $P$ . In particular, if

$$\Sigma_P = (\sigma_{ij}) \quad \text{where} \quad \sigma_{ij} := \int (s_i s_j)P(d\mathbf{s}) - \int s_i P(d\mathbf{s}) \cdot \int s_j P(d\mathbf{s}) \quad (5.16)$$

then  $\Sigma_P = \text{var}[\mathbf{x}]$ .

### 5.1.4 Independence

A collection of  $N$  random variables  $x_1, \dots, x_N$  is said to be **independent** if

$$\mathbb{P} \bigcap_{n=1}^N \{x_n \in B_n\} = \prod_{n=1}^N \mathbb{P}\{x_n \in B_n\} \quad (5.17)$$

for any  $B_1, \dots, B_N$ , where each  $B_n$  is a Borel subset of  $\mathbb{R}$ . In view of the definition of independence of events on page 87, the random variables  $x_1, \dots, x_N$  are independent precisely when sets of the form  $\{x_1 \in B_1\}, \dots, \{x_N \in B_N\}$  are independent events.

An infinite set of random variables  $\{x_n\}_{n=1}^\infty$  is called independent if any finite subset of  $\{x_n\}_{n=1}^\infty$  is independent in the sense of the preceding definition. We use the abbreviation **IID** for collections of random variables that are both independent and identically distributed. If we state that  $\{x_n\}_{n=1}^\infty$  are **IID copies of  $x$** , we mean that the random variables  $\{x_n\}_{n=1}^\infty$  are IID with the same distribution as  $x$ .

There are a number of equivalent ways that we can rephrase the definition of independence. Since  $\bigcap_{n=1}^N \{x_n \in B_n\} = \{(x_1, \dots, x_N) \in B_1 \times \dots \times B_N\}$ , a direct translation of (5.17) using (5.7) and (5.10) gives

$$P(B_1 \times \dots \times B_N) = \prod_{n=1}^N P_n(B_n) \quad (5.18)$$

Here  $P$  is the joint distribution of  $\mathbf{x} = (x_1, \dots, x_N)$  and  $P_n$  is its  $n$ th marginal. Thus, the elements of a random vector are independent if and only if their joint distribution equals the product distribution formed from their marginals (see fact 5.1.4).

The left- and right-hand sides of (5.18) agree for all  $B_1, \dots, B_N$  if they agree on certain families of sets, like the rectangles discussed in §5.1.2. For this reason a necessary and sufficient condition for independence of  $x_1, \dots, x_N$  is

$$F(s_1, \dots, s_N) = \prod_{n=1}^N F_n(s_n) \quad (5.19)$$

for all  $(s_1, \dots, s_N) \in \mathbb{R}^N$ , where  $F$  is the CDF of  $\mathbf{x}$  and  $F_1, \dots, F_N$  are the marginal CDFs. If the distribution of  $\mathbf{x}$  is absolutely continuous we can also test independence via its density:

**Fact 5.1.8** If  $\mathbf{x} = (x_1, \dots, x_N)$  has joint density  $p$  and marginals  $p_1, \dots, p_N$ , then  $x_1, \dots, x_N$  are independent if and only if

$$p(s_1, \dots, s_N) = \prod_{n=1}^N p_n(s_n) \quad \text{for all } (s_1, \dots, s_N) \in \mathbb{R}^N$$

**Example 5.1.4** Let  $\mathcal{L}(\mathbf{x}) = \mathcal{L}(x_1, \dots, x_N) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Suppose in addition that  $\boldsymbol{\Sigma}$  is diagonal, with  $n$ th diagonal component  $\sigma_n > 0$ . Then  $x_1, \dots, x_N$  are independent. To see this observe that for any  $\mathbf{s} = (s_1, \dots, s_N) \in \mathbb{R}^N$  we have

$$\begin{aligned} p(\mathbf{s}) &= (2\pi)^{-N/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{s} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{s} - \boldsymbol{\mu}) \right\} \\ &= \frac{1}{(2\pi)^{N/2} \prod_{n=1}^N \sigma_n} \exp \left\{ -\frac{1}{2} \sum_{n=1}^N (s_n - \mu_n)^2 \sigma_n^{-2} \right\} \end{aligned}$$

where the computation of the determinant and inverse of  $\boldsymbol{\Sigma}$  used facts 3.2.1 and 3.2.2 (see page 53). The last expression can be factored further, leading to

$$p(\mathbf{s}) = \prod_{n=1}^N \frac{1}{(2\pi)^{1/2} \sigma_n} \exp \left\{ -\frac{(s_n - \mu_n)^2}{2\sigma_n^2} \right\} = \prod_{n=1}^N p_n(s_n)$$

where  $p_n$  is the density of  $N(\mu_n, \sigma_n^2)$ . Independence now follows from fact 5.1.8.

The next fact is true in general but most easily verified in the density case, where you can apply (5.15) and fact 5.1.8.

**Fact 5.1.9** If  $x_1, \dots, x_N$  are independent and each  $x_n$  is integrable, then

$$\mathbb{E} \left[ \prod_{n=1}^N x_n \right] = \prod_{n=1}^N \mathbb{E}[x_n]$$

The definition of independence of random vectors is almost identical to the scalar case. Random vectors  $\mathbf{x}_1, \dots, \mathbf{x}_N$  in  $\mathbb{R}^K$  are called **independent** if

$$\mathbb{P} \bigcap_{n=1}^N \{\mathbf{x}_n \in B_n\} = \prod_{n=1}^N \mathbb{P}\{\mathbf{x}_n \in B_n\} \quad (5.20)$$

for any  $B_1, \dots, B_N$ , where each  $B_n$  is a Borel subset of  $\mathbb{R}^K$ .

**Fact 5.1.10** If  $\mathbf{x}_1, \dots, \mathbf{x}_N$  are independent random vectors in  $\mathbb{R}^K$  and  $f_1, \dots, f_N$  are any  $\mathcal{B}$ -measurable functions, then  $f_1(\mathbf{x}_1), \dots, f_N(\mathbf{x}_N)$  are also independent.

To see why, observe that  $f_n(\mathbf{x}_n) \in B_n$  if and only if  $\mathbf{x}_n \in f^{-1}(B_n)$ , as follows from the definition of preimage of a set (see 15.2). This leads to

$$\bigcap_{n=1}^N \{f_n(\mathbf{x}_n) \in B_n\} = \bigcap_{n=1}^N \{\mathbf{x}_n \in f^{-1}(B_n)\}$$

Applying independence of  $\mathbf{x}_1, \dots, \mathbf{x}_N$  now gives

$$\mathbb{P} \bigcap_{n=1}^N \{f_n(\mathbf{x}_n) \in B_n\} = \prod_{n=1}^N \mathbb{P}\{\mathbf{x}_n \in f^{-1}(B_n)\} = \prod_{n=1}^N \mathbb{P}\{f_n(\mathbf{x}_n) \in B_n\}$$

This completes the proof of fact 5.1.10.

**Fact 5.1.11** If  $\mathbf{x}$  and  $\mathbf{y}$  are independent, then  $\text{cov}(\mathbf{x}, \mathbf{y}) = 0$ .

Note that the converse is not true: one can construct examples of dependent random variables with zero covariance. However,

**Fact 5.1.12** If  $\mathbf{x}$  is multivariate Gaussian and  $\mathbf{A}$  and  $\mathbf{B}$  are conformable constant matrices, then  $\mathbf{Ax}$  and  $\mathbf{Bx}$  are independent if and only if  $\text{cov}(\mathbf{Ax}, \mathbf{Bx}) = 0$ .

We already proved a simple version of this in example 5.1.4. A full proof can be put together using the theory of characteristic functions (see, e.g., ch. 5 of Gut 2009). Below are two useful related results. The first one links orthogonal projection and multivariate Gaussians (ex. 5.4.11 requests a proof).

**Fact 5.1.13** Let  $S$  be any linear subspace of  $\mathbb{R}^N$ , let  $\mathbf{P} := \text{proj } S$  and let  $\mathbf{M}$  be the residual projection. If  $\mathcal{L}(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$  in  $\mathbb{R}^N$  for some  $\sigma^2 > 0$ , then  $\mathbf{Pz}$  and  $\mathbf{Mz}$  are independent.

**Fact 5.1.14** If  $w_1, \dots, w_N$  are independent with  $\mathcal{L}(w_n) = \mathcal{N}(\mu_n, \sigma_n^2)$  for all  $n$ , then

$$\mathcal{L} \left[ \alpha_0 + \sum_{n=1}^N \alpha_n w_n \right] = \mathcal{N} \left( \alpha_0 + \sum_{n=1}^N \alpha_n \mu_n, \sum_{n=1}^N \alpha_n^2 \sigma_n^2 \right)$$

This last fact is important because sums of arbitrary normals are not always normal. What is required is a multivariate normal distribution. Fact 5.1.14 works because independence pins down the joint distribution as multivariate normal. In particular,  $\mathcal{L}(w_1, \dots, w_N) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  where  $\mathbf{e}_n^\top \boldsymbol{\mu} = \mu_n$  and  $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$ . The expression for the variance in fact 5.1.14 follows from fact 4.1.10 on page 98.

### 5.1.5 Copulas

A **copula**  $C$  on  $\mathbb{R}^N$  is a multivariate CDF (see page 128) supported on the unit hypercube  $[0, 1]^N$  with the property that all its marginals are uniform on  $[0, 1]$ . In other words,  $C$  is a function of the form

$$C(s_1, \dots, s_N) = \mathbb{P}\{u_1 \leq s_1, \dots, u_N \leq s_N\} \quad (5.21)$$

where  $0 \leq s_n \leq 1$  and  $\mathcal{L}(u_n) = U[0, 1]$  for all  $n$ . There are many possible copulas because, while each  $u_n$  has its marginal distribution pinned down, there are infinitely many ways to specify the joint distribution.

**Example 5.1.5** The function  $C(s_1, s_2) = s_1 s_2$  on  $[0, 1]^2$  is called the **independence copula**. The marginal distributions are  $C(s_1, 1) = s_1$  and  $C(1, s_2) = s_2$  as required. (These are CDFs for the  $U[0, 1]$  distribution.)

**Example 5.1.6** The **Gumbel copulas** are the class of functions on  $[0, 1]^2$  defined by

$$C(s_1, s_2) = \exp \left\{ - \left[ (-\ln s_1)^\theta + (-\ln s_2)^\theta \right]^{1/\theta} \right\} \quad (\theta \geq 1)$$

The **Clayton copulas** are given by

$$C(s_1, s_2) = \left\{ \max \left[ s_1^{-\theta} + s_2^{-\theta} - 1, 0 \right] \right\}^{-1/\theta} \quad (\theta \geq -1, \theta \neq 0)$$

Both of these belong to a general class called the **Archimedean copulas**.

One reason that copulas are useful is that we can take univariate CDFs  $F_1, \dots, F_N$  and a copula  $C$  and create a multivariate CDF on  $\mathbb{R}^N$  via

$$F(s_1, \dots, s_N) = C(F_1(s_1), \dots, F_N(s_N)) \quad (s_n \in \mathbb{R}, n = 1, \dots, N) \quad (5.22)$$

The benefit of creating a multivariate CDF using this two step procedure is that in specifying a multivariate distribution we can separate out specification of the marginals, which give probabilities for isolated coordinates, and specification of the joint distribution, which defines the interactions.

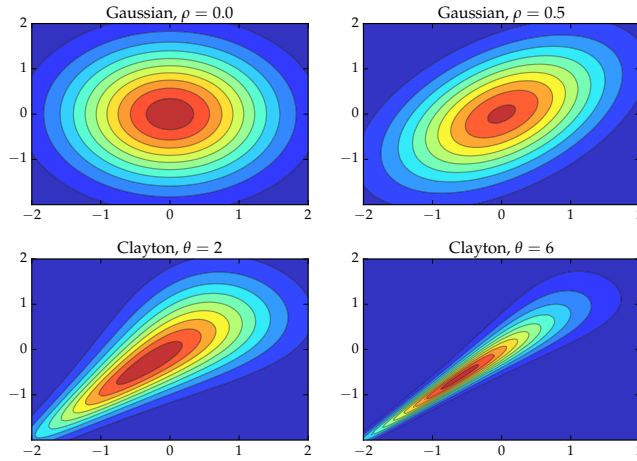


**Figure 5.5** Bivariate Gaussian (top) and non-Gaussian (bottom)

**Example 5.1.7** Bonhomme and Robin (2009) use copulas to model one component of earnings dynamics in a study based on three-year panels from the French Labor Force Survey. The cross sections are relatively large (around 30,000), allowing for flexible modeling of the marginal distributions via a mixture of normals. However, the time series dimension is short, so a one-parameter family of copulas is used to bind the marginals across time in a parsimonious way.

How rich is the set of multivariate CDFs we can create via (5.22)? It is universal. This is one part of a well-known theorem due to Sklar (1959):

**Theorem 5.1.1** *If  $F$  is any CDF on  $\mathbb{R}^N$  with marginals  $F_1, \dots, F_N$ , then there exists a copula  $C$  such that (5.22) holds. If each  $F_n$  is continuous, then this representation is unique.*

Copulas can be used to clarify a point raised in §5.1.3.2: normal marginals are not sufficient for the multivariate normal property. If  $F_1, \dots, F_N$  are univariate normal, then  $C(F_1(s_1), \dots, F_N(s_N))$  will equal the multivariate normal CDF for one choice of copula, called the Gaussian copula. Other choices lead to different distributions. Figure 5.5 gives an illustration where the top two distributions are multivariate normal and the bottom two are created by pairing the Clayton copula with univariate normal marginals. The distributions are represented as densities.

A related point is that while linear combinations of normal random variables are often thought to be normal, this isn't true unless the original random variables are multivariate normal (cf., fact 5.1.5 on page 133) or independent (which is a special case of multivariate normal). As an example, figure 5.6 shows a histogram of draws from

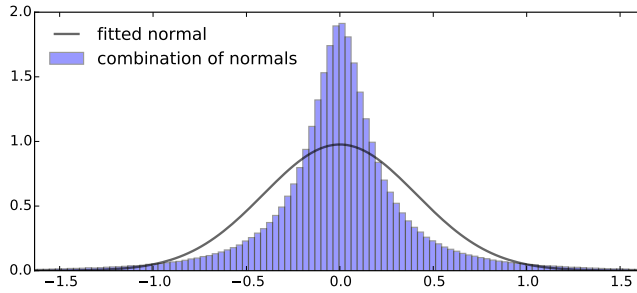


**Figure 5.6** Normality is not necessarily preserved under linear transformations

$w = x - y$  when  $(x, y)$  is distributed according to the density in the lower right panel of figure 5.5. For comparison, a normal density with the same mean and standard deviation is plotted.

### 5.1.6 Properties of Named Distributions

Here are some useful facts relating independence and certain common distributions.

**Fact 5.1.15** If  $x_1, \dots, x_N$  are independent and  $\mathcal{L}(x_n) = \chi^2(k_n)$ , then  $\mathcal{L}(\sum_n x_n) = \chi^2(\sum_n k_n)$ .

**Fact 5.1.16** If  $z$  and  $x$  are independent with  $\mathcal{L}(z) = \mathcal{N}(0, 1)$  and  $\mathcal{L}(x) = \chi^2(k)$ , then

$$z\sqrt{\frac{k}{x}} \text{ is } t \text{ distributed with } k \text{ degrees of freedom}$$

**Fact 5.1.17** If  $x_1$  and  $x_2$  are independent with  $\mathcal{L}(x_i) = \chi^2(k_i)$  for  $i = 1, 2$ , then

$$\frac{x_1/k_1}{x_2/k_2} \text{ is } F \text{ distributed with parameters } k_1, k_2$$

Let's note some facts connecting chi-squared and normal random variables. First, if  $\mathcal{L}(x) = \mathcal{N}(0, 1)$ , then  $\mathcal{L}(x^2) = \chi^2(1)$ . Exercise 5.4.4 asks you to prove this. Second, if a collection of standard normals are independent, then so are their squares (fact 5.1.10). Combining this result with fact 5.1.15 gives

**Fact 5.1.18** If  $\mathcal{L}(z_1, \dots, z_N) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ , then  $\mathcal{L}(\sum_{n=1}^N z_n^2) = \chi^2(N)$ .

The next result generalizes fact 5.1.18 and crops up frequently in statistics.

**Fact 5.1.19** If  $\mathcal{L}(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$  and  $\mathbf{A}$  is symmetric and idempotent, then

$$\mathcal{L}(\mathbf{z}^\top \mathbf{A} \mathbf{z}) = \chi^2(K) \quad \text{where} \quad K := \text{trace } \mathbf{A}$$

We can obtain fact 5.1.19 from fact 5.1.18 as follows. Since  $\mathbf{A}$  is symmetric, we use the spectral decomposition theorem (page 66) to write it as  $\mathbf{A} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top$ , where  $\mathbf{Q}$  is an orthogonal matrix and  $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_N)$  with  $\lambda_n :=$  the  $n$ th eigenvalue of  $\mathbf{A}$ . Setting  $\mathbf{y} := \mathbf{Q}^\top \mathbf{z}$ , we have

$$\mathbf{z}^\top \mathbf{A} \mathbf{z} = \mathbf{z}^\top \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top \mathbf{z} = \mathbf{y}^\top \mathbf{\Lambda} \mathbf{y} = \sum_{n=1}^N \lambda_n y_n^2$$

By exercise 3.5.35, idempotence of  $\mathbf{A}$  implies all eigenvalues lie in  $\{0, 1\}$ . By fact 3.2.7 on page 57, the number of nonzero eigenvalues is equal to  $\text{trace } \mathbf{A}$ . It follows that  $\mathbf{z}^\top \mathbf{A} \mathbf{z}$  is the sum of  $y_n^2$  over  $K := \text{trace } \mathbf{A}$  nonzero terms. By fact 5.1.18, a random variable with this property is  $\chi^2(K)$  whenever  $\mathbf{y}$  is standard normal. Combining  $\mathcal{L}(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$  and fact 5.1.5 on page 133, we have  $\mathcal{L}(\mathbf{y}) = \mathcal{N}(\mathbf{0}, \mathbf{Q}^\top \mathbf{Q})$ . But  $\mathbf{Q}$  is an orthogonal matrix, so  $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$ .

## 5.2 Conditioning and Expectation

Conditional expectation is one of the most important concepts in both economic theory and econometrics. This section gives a construction of expectation based around projection, which, in my view, is more intuitive and rewarding than both the elementary treatments using conditional distributions and the advanced treatments using Radon–Nykodym derivatives. The construction begins by formalizing the concepts of information and conditioning. We then build on our knowledge of orthogonal projection from §3.3 to frame conditional expectation as optimal prediction given limited information.

### 5.2.1 Conditional Distributions

Let's briefly discuss conditional distributions, focusing on the density case. Let  $x_1$  and  $x_2$  be random variables. The **conditional density** of  $x_2$  given  $x_1 = s_1$  is defined as

$$p(s_2 | s_1) := \frac{p(s_1, s_2)}{p(s_1)} \quad (5.23)$$

whenever the denominator is nonzero. Here  $p$  stands in for either joint, marginal, or conditional density, with the type determined by the argument. (For example,  $p(s_i)$  is

the marginal density of  $x_i$  evaluated at  $s_i$ .)

The law of total probability (see page 88) extends to the density case as follows: If  $(x_1, x_2)$  is a random vector on  $\mathbb{R}^2$ , then

$$p(s_2) = \int_{-\infty}^{\infty} p(s_2 | s_1) p(s_1) ds_1 \quad (s_2 \in \mathbb{R}) \quad (5.24)$$

To see this, fix  $s_2 \in \mathbb{R}$  and integrate the joint density to get the marginal, giving

$$p(s_2) = \int_{-\infty}^{\infty} p(s_1, s_2) ds_1$$

Combining this result with  $p(s_2 | s_1) = p(s_1, s_2) / p(s_1)$  yields (5.24). Bayes' law (see page 88) also extends to the density case, giving

$$p(s_2 | s_1) = \frac{p(s_1 | s_2) p(s_2)}{p(s_1)} \quad (5.25)$$

The same ideas extend to additional arguments in a natural way. For example, the conditional density of  $x_{k+1}, \dots, x_N$  given  $x_1 = s_1, \dots, x_k = s_k$  is defined by

$$p(s_{k+1}, \dots, s_N | s_1, \dots, s_k) = \frac{p(s_1, \dots, s_N)}{p(s_1, \dots, s_k)} \quad (5.26)$$

Rearranging this expression, we obtain a useful decomposition of the joint density:

$$p(s_1, \dots, s_N) = p(s_{k+1}, \dots, s_N | s_1, \dots, s_k) p(s_1, \dots, s_k) \quad (5.27)$$

## 5.2.2 The Space $L_2$

Suppose we want to predict random variable  $y$  using another variable  $x$ . In this case we should choose  $x$  such that  $x$  and  $y$  are expected to be close under most realizations of uncertainty. But what does "expected to be close" mean?

Perhaps the most common measure of distance between random variables is **mean squared error** (MSE), which is defined in this case as  $\mathbb{E}[(x - y)^2]$ . We met this concept previously in the linear prediction problem of §4.1.5. A common variation on MSE is **root mean squared error**, which, as the name suggests, is defined as

$$\|x - y\| := \sqrt{\mathbb{E}[(x - y)^2]} \quad (5.28)$$

On the right-hand side of (5.28) we have a measure of distance between random variables. On the left-hand side we have used the same symbol as we used for Euclidean

norm distance between two vectors. This is both deliberate and conventional, as there are many parallels between ordinary vector space with the Euclidean norm and the set of random variables combined with the “norm” defined in (5.28). Let’s now clarify and formalize these ideas.

The first geometric concept we defined for vectors was inner product. Analogously, we define here the **inner product between two random variables**  $x$  and  $y$  as

$$\langle x, y \rangle := \mathbb{E}[xy] \quad (5.29)$$

Of course we want this definition to make sense, and the Cauchy–Schwarz inequality for random variables (page 97) tells us that  $\mathbb{E}[xy]$  will be finite and well-defined whenever  $x$  and  $y$  both have finite second moments. Hence we restrict attention to this set, which is commonly denoted  $L_2$ . That is,

$$L_2 := \{ \text{all random variables } x \text{ on } (\Omega, \mathcal{F}, \mathbb{P}) \text{ with } \mathbb{E}[x^2] < \infty \}$$

Here’s a list of some of the ways that  $L_2$  inner product is similar to the Euclidean inner product on  $\mathbb{R}^N$ .

**Fact 5.2.1** For any  $\alpha, \beta \in \mathbb{R}$  and any  $x, y, z \in L_2$ , the following statements are true:

- (i)  $\langle x, y \rangle = \langle y, x \rangle$ .
- (ii)  $\langle \alpha x, \beta y \rangle = \alpha \beta \langle x, y \rangle$ .
- (iii)  $\langle x, \alpha y + \beta z \rangle = \alpha \langle x, y \rangle + \beta \langle x, z \rangle$ .

If you compare fact 5.2.1 with fact 2.1.1 on page 13 you’ll see that the correspondence is essentially one-to-one. The properties in fact 5.2.1 follow from the definition of the inner product and linearity of  $\mathbb{E}$  (see fact 4.1.5 on page 95).

Analogous to the Euclidean vector norm, we define the  **$L_2$  norm** by

$$\|x\| := \sqrt{\langle x, x \rangle} := \sqrt{\mathbb{E}[x^2]} \quad (x \in L_2) \quad (5.30)$$

From this norm we get a notion of distance  $\|x - y\|$  between random variables that agrees with the notion of root MSE in (5.28). As with inner product, the properties of the  $L_2$  norm and vector norm are largely identical. (Compare the following with fact 2.1.2 on page 13.)

**Fact 5.2.2** For any  $\alpha \in \mathbb{R}$  and any  $x, y \in L_2$ , the following statements are true:

- (i)  $\|x\| \geq 0$  and  $\|x\| = 0$  if and only if  $x = 0$
- (ii)  $\|\alpha x\| = |\alpha| \|x\|$
- (iii)  $\|x + y\| \leq \|x\| + \|y\|$

$$(iv) \quad |\langle x, y \rangle| \leq \|x\| \|y\|$$

Point (i) requires a clarification, which is given below. Property (ii) is immediate from (5.30) and the linearity of  $\mathbb{E}$ . Property (iii) is called the **triangle inequality**, as in the vector case. Property (iv) is just the **Cauchy–Schwarz inequality** from page 97, restated in new notation. As in the vector case, the triangle inequality can be proved from the Cauchy–Schwarz inequality (see page 13). For a proof, see exercise 5.4.15.

Returning to (i), it isn't true that  $\|x\| = 0$  implies  $x(\omega) = 0$  for all  $\omega \in \Omega$ . What we can say is that if  $\|x\| = 0$  then  $\mathbb{P}\{x = 0\} = 1$  (see, e.g., Williams 1991, §5.2). If we agree not to distinguish between random variables that only differ with zero probability, then (i) is valid. This is the standard convention when dealing with  $L_2$ .

### 5.2.2.1 Linear Subspaces in $L_2$

From (ii) and (iii) of fact 5.2.2, we see that if  $\alpha, \beta \in \mathbb{R}$  and  $x, y \in L_2$ , then  $\alpha x + \beta y$  also has finite norm, and is therefore in  $L_2$ . More generally, any **linear combination**

$$\alpha_1 x_1 + \cdots + \alpha_K x_K, \quad \alpha_k \in \mathbb{R}, \quad x_k \in L_2$$

is again in  $L_2$ . As in the vector case (see §2.1.2), when  $X$  is a subset of  $L_2$ , the set of finite linear combinations that can be formed from elements of  $X$  is called the **span** of  $X$ , and denoted by  $\text{span } X$ .

**Example 5.2.1** If  $x \in L_2$  and  $\mathbb{1} := \mathbb{1}_\Omega$  is the constant random variable always equal to 1, then  $\text{span}\{\mathbb{1}, x\}$  is the set of random variables

$$\alpha + \beta x := \alpha \mathbb{1} + \beta x \quad \text{for scalars } \alpha, \beta \quad (5.31)$$

This is the set  $\mathcal{L}$  introduced on page 98 when we discussed best linear predictors.

A subset  $S$  of  $L_2$  is called a **linear subspace** of  $L_2$  if it is closed under addition and scalar multiplication. That is, for each  $x, y \in S$  and  $\alpha, \beta \in \mathbb{R}$ , we have  $\alpha x + \beta y \in S$ .

**Example 5.2.2** The span of any set of elements of  $L_2$  is a linear subspace in  $L_2$ . (Compare with example 2.1.9 on page 20.)

**Example 5.2.3** The set  $Z := \{x \in L_2 : \mathbb{E}x = 0\}$  is a linear subspace of  $L_2$  because

$$x, y \in Z \text{ and } \alpha, \beta \in \mathbb{R} \implies \mathbb{E}[\alpha x + \beta y] = \alpha \mathbb{E}[x] + \beta \mathbb{E}[y] = 0$$

As in  $\mathbb{R}^N$ , an **orthonormal basis** of a linear subspace  $S$  of  $L_2$  is a set  $\{u_1, \dots, u_K\} \subset S$  with the property that

$$\langle u_j, u_k \rangle = \mathbb{1}\{j = k\} \quad \text{and} \quad \text{span}\{u_1, \dots, u_K\} = S$$

**Example 5.2.4** Let  $x \in L_2$  so that  $S := \text{span}\{\mathbb{1}, x\}$  is as in (5.31). If we define

$$u_1 := \mathbb{1} \quad \text{and} \quad u_2 := \frac{x - \mu}{\sigma_x}$$

Then

$$\langle u_1, u_2 \rangle = \mathbb{E}[u_1 u_2] = \mathbb{E}\left[\frac{x - \mu}{\sigma_x}\right] = 0$$

Clearly,  $\|u_1\| = \|u_2\| = 1$ , so this pair is orthonormal. It's also straightforward to show that  $\text{span}\{u_1, u_2\} = \text{span}\{\mathbb{1}, x\}$ , so  $\{u_1, u_2\}$  is an orthonormal basis for  $S$ .

### 5.2.3 Projections in $L_2$

Given that inner products and norms in  $L_2$  share so many properties with their Euclidean cousins, it's not surprising that many results about  $\mathbb{R}^N$  carry over to  $L_2$ . Often we can just lift the proofs across with barely a change to notation. We know a great deal about vectors and Euclidean space, so this process yields many insights about random variables. Let's illustrate this in the context of orthogonal projection.

As in the Euclidean case, if  $\langle x, y \rangle = 0$ , then we say that  $x$  and  $y$  are **orthogonal**, and write  $x \perp y$ . This terminology is used frequently in econometrics. Its importance stems partly from the next fact, the proof of which is a short exercise.

**Fact 5.2.3** If  $x, y \in L_2$  and  $\mathbb{E}x = 0$  or  $\mathbb{E}y = 0$  then  $x \perp y \iff \text{cov}[x, y] = 0$ .

Now consider the following problem. Given  $y \in L_2$  and linear subspace  $S \subset L_2$ , we seek the closest element  $\hat{y}$  of  $S$  to  $y$ . Closeness is in terms of  $L_2$  norm, so  $\hat{y}$  is the minimizer of  $\|y - z\|$  over all  $z \in S$ . That is, we seek

$$\hat{y} = \underset{z \in S}{\text{argmin}} \|y - z\| = \underset{z \in S}{\text{argmin}} \sqrt{\mathbb{E}[(y - z)^2]} \quad (5.32)$$

The next theorem mimics theorem 2.2.1 on page 29. A nice treatment that covers the current setting can be found in chapter 2 of Cheney (2001).

**Theorem 5.2.1** (Orthogonal Projection Theorem IV) *Let  $y \in L_2$  and let  $S$  be any nonempty closed linear subspace of  $L_2$ . The following statements are true:*

- (i) *The optimization problem (5.32) has exactly one solution.*
- (ii)  *$\hat{y} \in L_2$  is the unique solution<sup>3</sup> to (5.32) if and only if  $\hat{y} \in S$  and  $y - \hat{y} \perp S$ .*

---

3. Uniqueness is a probability one statement: Any two solutions  $\hat{y}_1$  and  $\hat{y}_2$  satisfy  $\mathbb{P}\{\hat{y}_1 = \hat{y}_2\} = 1$ . We've already agreed not to distinguish between such elements. See the discussion after fact 5.2.2.

The statement that  $S$  is closed means that  $\{x_n\} \subset S$  and  $x \in L_2$  with  $\|x_n - x\| \rightarrow 0$  implies  $x \in S$ . This condition is true for all the linear subspaces we want to work with.

Analogous with the case of  $\mathbb{R}^N$ , the random variable  $\hat{y}$  in theorem 5.2.1 is called the **orthogonal projection of  $y$  onto  $S$** . Holding  $S$  fixed, we can think of the operation

$$y \mapsto \text{the orthogonal projection of } y \text{ onto } S$$

as a function from  $L_2$  to  $L_2$ . The function will be denoted by  $\mathbf{P}$ , so that, for each  $y \in L_2$ , the symbol  $\mathbf{P}y$  represents the image of  $y$  under  $\mathbf{P}$ , which is the orthogonal projection  $\hat{y}$ . As before,  $\mathbf{P}$  is called the **orthogonal projection onto  $S$** , and we write  $\mathbf{P} = \text{proj } S$ . Since  $\mathbf{P}y$  solves (5.32), we can and do interpret  $\mathbf{P}y$  as the *best predictor of  $y$  from within the collection of random variables contained in  $S$* .

**Fact 5.2.4** Let  $S$  be any linear subspace of  $L_2$ , and let  $\mathbf{P} = \text{proj } S$ . The following statements are true:

(i)  $\mathbf{P}$  is a linear function.

Moreover, for any  $y \in L_2$ , we have

(ii)  $\mathbf{P}y \in S$ ,

(iii)  $y - \mathbf{P}y \perp S$ ,

(iv)  $\|y\|^2 = \|\mathbf{P}y\|^2 + \|y - \mathbf{P}y\|^2$ ,

(v)  $\|\mathbf{P}y\| \leq \|y\|$ , and

(vi)  $\mathbf{P}y = y$  if and only if  $y \in S$ .

Compare with theorem 2.2.2 on page 31. In (i), the statement that  $\mathbf{P}$  is linear means that  $\mathbf{P}(\alpha x + \beta y) = \alpha \mathbf{P}x + \beta \mathbf{P}y$  for all  $x, y \in L_2$  and  $\alpha, \beta \in \mathbb{R}$ .

The next two results are  $L_2$  versions of fact 2.2.7 on page 32 and fact 2.2.6 on page 32 respectively. The proofs are essentially the same.

**Fact 5.2.5** Let  $S_i$  be a linear subspace of  $L_2$  for  $i = 1, 2$  and let  $\mathbf{P}_i = \text{proj } S_i$ . If  $S_1 \subset S_2$ , then  $\mathbf{P}_1 \mathbf{P}_2 y = \mathbf{P}_1 y$  for all  $y \in L_2$ .

**Fact 5.2.6** If  $\{u_1, \dots, u_K\}$  is an orthonormal basis of  $S$ , then, for all  $y \in L_2$ ,

$$\mathbf{P}y = \sum_{k=1}^K \langle y, u_k \rangle u_k \quad (5.33)$$

**Example 5.2.5** The mean of a random variable  $x$  can be thought of as the “best predictor of  $x$  within the set of constants.” To see this from a new perspective, let  $S := \text{span}\{\mathbb{1}\}$ , where  $\mathbb{1} := \mathbb{1}_\Omega$ , and let  $\mathbf{P} := \text{proj } S$ . The object  $\mathbf{P}x$  is precisely the best predictor of  $x$  within the class of constant random variables. Not surprisingly,  $\mathbf{P}x = \mu \mathbb{1}$ ,

where  $\mu := \mathbb{E} x$ . The easiest way to check this is to observe that  $\{\mathbb{1}\}$  is an orthonormal set spanning  $S$ , and hence, by (5.33),

$$\mathbf{P}x = \langle x, \mathbb{1} \rangle \mathbb{1} = \mathbb{E}[x\mathbb{1}]\mathbb{1} = \mathbb{E}[x]\mathbb{1} = \mu \mathbb{1}$$

You can also check the claim that  $\mu \mathbb{1}$  is the projection of  $x$  onto  $S$  by verifying the conditions in (ii) of theorem 5.2.1.

**Example 5.2.6** Let's revisit the problem of best linear predictors from §4.1.5, establishing the same result in a different way. Fix  $x, y \in L_2$  and consider projecting  $y$  onto  $S := \text{span}\{\mathbb{1}, x\}$ . The set  $S$  is as given in (5.31). The problem of projecting  $y$  onto  $S$  is equivalent to the best linear prediction problem in (4.21) on page 99. To implement the projection recall from example 5.2.4 that

$$u_1 := \mathbb{1} \quad \text{and} \quad u_2 := \frac{x - \mu}{\sigma_x}$$

form an orthonormal basis for  $S$ . Hence, letting  $\mathbf{P} = \text{proj } S$  and applying (5.33),

$$\mathbf{P}y = \langle y, u_1 \rangle u_1 + \langle y, u_2 \rangle u_2 = \mathbb{E}[y] + \frac{\text{cov}[x, y]}{\text{var}[x]}(x - \mathbb{E}[x])$$

Alternatively, we can write

$$\mathbf{P}y = \alpha^* + \beta^* x \quad \text{where} \quad \beta^* := \frac{\text{cov}[x, y]}{\text{var}[x]} \quad \text{and} \quad \alpha^* := \mathbb{E}[y] - \beta^* \mathbb{E}[x]$$

These are the same solutions we obtained using calculus in §4.1.5.

### 5.2.3.1 Population Regression

Let's consider an extension of the best linear prediction problem of example 5.2.6 to a more general setting, where the information for predicting  $y$  is a random vector  $\mathbf{x}$  in  $\mathbb{R}^K$ . As before, prediction will be linear, with the aim of minimizing squared error. Hence the object we seek is the  $L_2$  orthogonal projection of  $y$  onto the linear subspace

$$\text{span}\{\mathbf{x}\} := \text{all random variables of the form } \mathbf{x}^\top \mathbf{b} \text{ for some } \mathbf{b} \in \mathbb{R}^K$$

To make sure that  $\text{span}\{\mathbf{x}\} \subset L_2$ , we will assume that  $\mathbb{E}[\mathbf{x}^\top \mathbf{x}] < \infty$ .

**Fact 5.2.7** If  $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$  is positive definite, then the projection  $\mathbf{P}y$  of any  $y \in L_2$  onto  $\text{span}\{\mathbf{x}\}$  is given by

$$\hat{y} = \mathbf{x}^\top \mathbf{b}^* \quad \text{where} \quad \mathbf{b}^* := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]^{-1} \mathbb{E}[\mathbf{x}y] \quad (5.34)$$

Exercise 5.4.16 asks you to prove this. Positive definiteness of  $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$  ensures invertibility, without which the vector  $\mathbf{b}^*$  is not uniquely defined (see ex. 3.5.34). Note that, by the definition of orthogonal projections,  $\mathbf{b}^*$  necessarily satisfies

$$\mathbf{b}^* = \underset{\mathbf{a} \in \mathbb{R}^K}{\operatorname{argmin}} \mathbb{E}[(y - \mathbf{x}^\top \mathbf{a})^2]$$

The linear prediction problem considered here is also called **population linear regression**. The “population” qualifier is because we are using the true joint distribution of  $(\mathbf{x}, y)$  when we compute expectations. Population regression has a sample counterpart called multivariate linear regression, based on observations of  $(\mathbf{x}, y)$ . We will discuss it at length in chapter 11.

## 5.2.4 Measurability

While the use of projections to obtain best linear predictors in example 5.2.6 and fact 5.2.7 is elegant and useful, we don’t always want to constrain ourselves to linear predictions. To drop the linearity requirement, we change the linear subspaces used for projection from the set of linear functions of  $\mathbf{x}$  to the set of arbitrary functions of  $\mathbf{x}$ . As we now show, the best predictor that results is exactly the conditional expectation with respect to  $\mathbf{x}$ .

The key to this process is defining the subspace to project onto. In the language of probability, the subspace of arbitrary real-valued functions of  $\mathbf{x}$  is called the  $\mathbf{x}$ -measurable functions. Let’s take these ideas one step at a time, starting with the notion of measurability with respect to information sets.

Let  $\mathcal{G} := \{x_1, \dots, x_D\}$  be any set of random variables and let  $z$  be any other random variable. The variable  $z$  is said to be  **$\mathcal{G}$ -measurable** if there exists a  $\mathcal{B}$ -measurable function  $g: \mathbb{R}^D \rightarrow \mathbb{R}$  such that

$$z = g(x_1, \dots, x_D) \tag{5.35}$$

As usual, this equality between random variables should be interpreted pointwise, so the more explicit version is

$$z(\omega) = g(x_1(\omega), \dots, x_D(\omega)) \quad \text{for all } \omega \in \Omega \tag{5.36}$$

In this context,  $\mathcal{G}$  is sometimes referred to as the **information set**. When convenient we’ll also write  $\mathbf{x} = (x_1, \dots, x_D)$  and say that  $z$  is  $\mathbf{x}$ -measurable. Similar terminology will be used for scalars and matrices. For example, if  $\mathbf{X}$  is a random matrix, then  $\mathbf{X}$ -measurability means  $\mathcal{G}$ -measurability when  $\mathcal{G}$  lists all elements of  $\mathbf{X}$ .

The intuitive meaning of  $\mathcal{G}$ -measurability of  $z$  is that  $z$  is completely determined

by the elements in  $\mathcal{G}$ . Imagine it like this: When uncertainty is realized, some  $\omega$  is selected from  $\Omega$ . We don't get to view  $\omega$  itself, but we do get to observe the realized values  $x_1(\omega), \dots, x_D(\omega)$  on the right-hand side of (5.36). If  $z$  is  $\mathcal{G}$ -measurable, we can now calculate the realized value  $z(\omega)$  of  $z$ , even without knowing  $\omega$ .

**Example 5.2.7** Let  $x, y$  and  $z$  be random variables and let  $\alpha$  and  $\beta$  be scalars. If  $z = \alpha x + \beta y$ , then  $z$  is  $\{x, y\}$ -measurable (take  $g(s, t) := \alpha s + \beta t$ ).

**Example 5.2.8** If  $x_1, \dots, x_N$  are random variables and  $\mathcal{G} := \{x_1, \dots, x_N\}$ , then the sample mean  $\bar{x}_N := \frac{1}{N} \sum_{n=1}^N x_n$  is  $\mathcal{G}$ -measurable.

**Example 5.2.9** Let  $x$  and  $y$  be independent and nondegenerate. Then  $y$  is not  $x$ -measurable, for if it were, then we would have  $y = g(x)$  for some function  $g$ , contradicting independence of  $x$  and  $y$ .<sup>4</sup>

**Example 5.2.10** Let  $y$  be a noisy signal of  $x$ , in the sense that  $y = x + u$  where  $x$  and  $u$  are independent and  $u$  is nondegenerate. Then  $y$  is not  $x$ -measurable. Intuitively, even if we know the value of  $x$ , the realization of  $y$  cannot be computed until we know the value of the disturbance term  $u$ . More formally, if  $y$  is  $x$ -measurable, then  $y = g(x)$  for some  $g$ , in which case  $u = g(x) - x$ . But then  $u$  is  $x$ -measurable, contradicting independence.

**Example 5.2.11** Let  $y = \alpha$ , where  $\alpha$  is a constant. This degenerate random variable is  $\mathcal{G}$ -measurable for any information set  $\mathcal{G}$ , because  $y$  is already deterministic. For example, if  $\mathcal{G} = \{x_1, \dots, x_p\}$ , then we can take  $y = g(x_1, \dots, x_p) = \alpha + \sum_{i=1}^p 0x_i$ .

If  $x$  and  $y$  are known given the information in  $\mathcal{G}$ , then a third random variable that depends only on  $x$  and  $y$  is likewise known given  $\mathcal{G}$ . Hence  $\mathcal{G}$ -measurability is preserved under the taking of sums, products, and so on. In particular,

**Fact 5.2.8** Let  $\alpha, \beta$  be any scalars, and let  $x$  and  $y$  be random variables. If  $x$  and  $y$  are both  $\mathcal{G}$ -measurable, then  $u := xy$  and  $v := \alpha x + \beta y$  are also  $\mathcal{G}$ -measurable.

Suppose now that  $\mathcal{G} \subset L_2$ , which just means that all the random variables in  $\mathcal{G}$  have finite second moment. Consider the set

$$L_2(\mathcal{G}) := \{\text{all } \mathcal{G}\text{-measurable random variables in } L_2\}$$

In view of fact 5.2.8, we have the following result:

**Fact 5.2.9** For any  $\mathcal{G} \subset L_2$ , the set  $L_2(\mathcal{G})$  is a linear subspace of  $L_2$ .

---

4. A tighter proof can be given using the language of  $\sigma$ -algebras, although the intuition is less immediate.

This furnishes us with a subspace to project onto, allowing us to define conditional expectations. But before we do so let's cover ordering of information sets by set inclusion.

**Fact 5.2.10** If  $\mathcal{G} \subset \mathcal{H}$  and  $z$  is  $\mathcal{G}$ -measurable, then  $z$  is  $\mathcal{H}$ -measurable.

This is intuitive: If  $z$  is known once the variables in  $\mathcal{G}$  are known, then it is certainly known when the extra information provided by  $\mathcal{H}$  is available.

**Example 5.2.12** Let  $x_1, x_2$  and  $y$  be random variables and let

$$\mathcal{G} := \{x_1\} \subset \{x_1, x_2\} =: \mathcal{H}$$

If  $y$  is  $\mathcal{G}$ -measurable, then  $y = g(x_1)$  for some  $\mathcal{B}$ -measurable  $g$ . But then  $y$  will also be  $\mathcal{H}$ -measurable. For example, we can write  $y = h(x_1, x_2)$  where  $h(x_1, x_2) = g(x_1) + 0x_2$ .

The next result is immediate from fact 5.2.10.

**Fact 5.2.11** If  $\mathcal{G} \subset \mathcal{H}$ , then  $L_2(\mathcal{G}) \subset L_2(\mathcal{H})$ .

Incidentally, most advanced probability texts give a definition of  $\mathcal{G}$ -measurability framed in terms of  $\sigma$ -algebras. This is necessary if you wish to work through all related proofs. But the definition given here is equivalent and intuitive. See theorem 4.4 of Çınlar (2011) if you wish to learn more.

### 5.2.5 Conditional Expectation

Now it's time to define conditional expectations. Let  $\mathcal{G} \subset L_2$  and  $y$  be some random variable in  $L_2$ . The **conditional expectation** of  $y$  given  $\mathcal{G}$  is written as  $\mathbb{E}[y | \mathcal{G}]$  or  $\mathbb{E}^{\mathcal{G}}[y]$ , and defined as the closest  $\mathcal{G}$ -measurable random variable to  $y$ . More formally,

$$\mathbb{E}[y | \mathcal{G}] := \operatorname{argmin}_{z \in L_2(\mathcal{G})} \|y - z\| \quad (5.37)$$

This definition makes sense. Our understanding of the conditional expectation  $\mathbb{E}[y | \mathcal{G}]$  is that it is the best predictor of  $y$  given the information contained in  $\mathcal{G}$ . The definition in (5.37) says the same thing. It simultaneously restricts  $\mathbb{E}[y | \mathcal{G}]$  to be  $\mathcal{G}$ -measurable, so we can actually compute it once the variables in  $\mathcal{G}$  are realized, and selects  $\mathbb{E}[y | \mathcal{G}]$  as the closest such variable to  $y$  in terms of RMSE.

But does the minimizer generally exist? And is it unique? The answer to both questions is yes. Comparing (5.37) and (5.32), we see that

$$\mathbb{E}[y | \mathcal{G}] = \mathbf{P}y \quad \text{when } \mathbf{P} := \operatorname{proj} L_2(\mathcal{G})$$

By theorem 5.2.1, the projection exists and is unique. Theorem 5.2.1 also characterizes  $\mathbb{E}[y | \mathcal{G}]$  as the unique point  $\hat{y}$  in  $L_2$  such that

$$\hat{y} \in L_2(\mathcal{G}) \quad \text{and} \quad y - \hat{y} \perp z \quad \text{for all } z \in L_2(\mathcal{G})$$

Rewriting these conditions in a slightly different way, we can give an alternative (and equivalent) definition of conditional expectation:  $\hat{y} \in L_2$  is the **conditional expectation** of  $y$  given  $\mathcal{G}$  if

- (i)  $\hat{y}$  is  $\mathcal{G}$ -measurable and
- (ii)  $\mathbb{E}[\hat{y}z] = \mathbb{E}[yz]$  for all  $\mathcal{G}$ -measurable  $z \in L_2$ .

When convenient we'll also use symbols like  $\mathbb{E}[y | x_1, \dots, x_D]$  or  $\mathbb{E}[y | \mathbf{x}]$ . These are the same as  $\mathbb{E}[y | \mathcal{G}]$  when  $\mathcal{G}$  is defined as the information set containing the variables we condition on.

**Example 5.2.13** If  $x$  and  $u$  are independent,  $\mathbb{E}u = 0$  and  $y = x + u$ , then  $\mathbb{E}[y | x] = x$ . To prove this we need to show that  $x$  satisfies (i)–(ii) above. Clearly,  $x$  is  $x$ -measurable. For (ii) we need to show that  $\mathbb{E}[xz] = \mathbb{E}[yz]$  for all  $x$ -measurable  $z$ . This translates to the claim that

$$\mathbb{E}[xg(x)] = \mathbb{E}[(x + u)g(x)]$$

for any  $\mathcal{B}$ -measurable  $g$ , which is true from independence and  $\mathbb{E}u = 0$ .

**Fact 5.2.12** Given  $\mathbf{x} \in \mathbb{R}^D$  and  $y$  in  $L_2$ , there exists a  $\mathcal{B}$ -measurable function  $f^*: \mathbb{R}^D \rightarrow \mathbb{R}$  such that  $\mathbb{E}[y | \mathbf{x}] = f^*(\mathbf{x})$ .

This is obvious because  $\mathbb{E}[y | \mathbf{x}]$  is by definition  $\mathbf{x}$ -measurable. A prediction of  $y$  given  $\mathbf{x}$  has to be a deterministic function of  $\mathbf{x}$ . The particular function  $f^*$  satisfying  $f^*(\mathbf{x}) = \mathbb{E}[y | \mathbf{x}]$  is called the **regression function** of  $y$  given  $\mathbf{x}$ . It is, by construction, the best predictor of  $y$  given  $\mathbf{x}$  in terms of mean squared error.

The next example shows that when  $x$  and  $y$  are linked by a conditional density, then our definition of conditional expectation reduces to the one seen in elementary probability texts. The proof of the claim in the example is the topic of exercise 5.4.23.

**Example 5.2.14** If  $x$  and  $y$  are random variables and  $p(y | x)$  is the conditional density of  $y$  given  $x$ , then

$$\mathbb{E}[y | x] = \int t p(t | x) dt$$

Here are some goodies we can harvest using the fact that  $\mathbb{E}[\cdot | \mathcal{G}] = \text{proj } L_2(\mathcal{G})$ .

**Fact 5.2.13** Let  $x$  and  $y$  be random variables in  $L_2$ , let  $\alpha$  and  $\beta$  be scalars, and let  $\mathcal{G}$  and  $\mathcal{H}$  be subsets of  $L_2$ . The following properties hold:

- (i) Linearity:  $\mathbb{E}[\alpha x + \beta y | \mathcal{G}] = \alpha \mathbb{E}[x | \mathcal{G}] + \beta \mathbb{E}[y | \mathcal{G}]$ .
- (ii) If  $\mathcal{G} \subset \mathcal{H}$ , then  $\mathbb{E}[\mathbb{E}[y | \mathcal{H}] | \mathcal{G}] = \mathbb{E}[y | \mathcal{G}]$  and  $\mathbb{E}[\mathbb{E}[y | \mathcal{G}]] = \mathbb{E}[y]$ .
- (iii) If  $y$  is independent of the variables in  $\mathcal{G}$ , then  $\mathbb{E}[y | \mathcal{G}] = \mathbb{E}[y]$ .
- (iv) If  $y$  is  $\mathcal{G}$ -measurable, then  $\mathbb{E}[y | \mathcal{G}] = y$ .
- (v) If  $x$  is  $\mathcal{G}$ -measurable, then  $\mathbb{E}[xy | \mathcal{G}] = x \mathbb{E}[y | \mathcal{G}]$ .

Some of these claims are given as solved exercises. Others follow directly from the definition and from the properties of projections. For example, in (iv) we are saying that if  $y \in L_2(\mathcal{G})$ , then  $y$  is projected into itself. This is immediate from (vi) in fact 5.2.4 on page 146.

Part (ii) is called the **the law of iterated expectations**. The law follows from the property of orthogonal projections given in fact 2.2.7 on page 32: Projecting onto the bigger subspace  $L_2(\mathcal{H})$  and from there onto  $L_2(\mathcal{G})$  is the same as projecting directly onto the smaller subspace  $L_2(\mathcal{G})$ .

Part (v) is called **conditional determinism**. It says that if  $x$  is known given  $\mathcal{G}$  then we can treat it as a constant in the conditional expectation.

### 5.2.5.1 Conditional Expectations and Prediction

To recap, given  $y \in L_2$  and random vector  $\mathbf{x}$  in  $\mathbb{R}^D$ , the conditional expectation  $\mathbb{E}[y | \mathbf{x}]$  is a function  $f^*$  of  $\mathbf{x}$ , called the regression function of  $y$  given  $\mathbf{x}$ , such that  $f^*(\mathbf{x})$  is the best predictor of  $y$  in terms of MSE. In other words,

$$f^* = \operatorname{argmin}_{g \in G} \mathbb{E}[(y - g(\mathbf{x}))^2] \quad (5.38)$$

where  $G$  is the functions from  $\mathbb{R}^D$  to  $\mathbb{R}$  with  $g(\mathbf{x}) \in L_2$ . In fact a more explicit result can be obtained: For any  $g \in G$ , we have

$$\mathbb{E}[(y - g(\mathbf{x}))^2] = \mathbb{E}[(y - f^*(\mathbf{x}))^2] + \mathbb{E}[(f^*(\mathbf{x}) - g(\mathbf{x}))^2] \quad (5.39)$$

This implies (5.38) because  $(f^*(\mathbf{x}) - g(\mathbf{x}))^2 \geq 0$ .

To prove (5.39), let  $f^*$  be the regression function, pick any  $g \in G$  and observe that

$$\begin{aligned} (y - g(\mathbf{x}))^2 &= (y - f^*(\mathbf{x}) + f^*(\mathbf{x}) - g(\mathbf{x}))^2 \\ &= (y - f^*(\mathbf{x}))^2 + 2(y - f^*(\mathbf{x}))(f^*(\mathbf{x}) - g(\mathbf{x})) + (f^*(\mathbf{x}) - g(\mathbf{x}))^2 \end{aligned}$$

Consider the expectation of the cross-product term. From the law of iterated expectations,

$$\mathbb{E}\{(y - f^*(\mathbf{x}))(f^*(\mathbf{x}) - g(\mathbf{x}))\} = \mathbb{E}\{\mathbb{E}[(y - f^*(\mathbf{x}))(f^*(\mathbf{x}) - g(\mathbf{x})) | \mathbf{x}]\} \quad (5.40)$$

Using conditional determinism we can rewrite the term inside the curly brackets on the right-hand side of (5.40) as  $(f^*(\mathbf{x}) - g(\mathbf{x}))\mathbb{E}[(y - f^*(\mathbf{x})) | \mathbf{x}]$ . For the second term in this product we have

$$\mathbb{E}[y - f^*(\mathbf{x}) | \mathbf{x}] = \mathbb{E}[y | \mathbf{x}] - \mathbb{E}[f^*(\mathbf{x}) | \mathbf{x}] = \mathbb{E}[y | \mathbf{x}] - f^*(\mathbf{x}) = 0$$

Hence the expectation in (5.40) is zero. Equation (5.39) now follows.

### 5.2.6 The Vector Case

Conditional expectations of random matrices are defined using the notion of conditional expectations for scalar random variables. For example, given random matrices  $\mathbf{X}$  and  $\mathbf{Y}$ , we set

$$\mathbb{E}[\mathbf{Y} | \mathbf{X}] := \begin{pmatrix} \mathbb{E}[y_{11} | \mathbf{X}] & \cdots & \mathbb{E}[y_{1K} | \mathbf{X}] \\ \vdots & & \vdots \\ \mathbb{E}[y_{N1} | \mathbf{X}] & \cdots & \mathbb{E}[y_{NK} | \mathbf{X}] \end{pmatrix}$$

We also define

- (i)  $\text{cov}[\mathbf{x}, \mathbf{y} | \mathbf{Z}] := \mathbb{E}[\mathbf{xy}^\top | \mathbf{Z}] - \mathbb{E}[\mathbf{x} | \mathbf{Z}]\mathbb{E}[\mathbf{y} | \mathbf{Z}]^\top$
- (ii)  $\text{var}[\mathbf{x} | \mathbf{Z}] := \mathbb{E}[\mathbf{xx}^\top | \mathbf{Z}] - \mathbb{E}[\mathbf{x} | \mathbf{Z}]\mathbb{E}[\mathbf{x} | \mathbf{Z}]^\top$

Using the definitions, one can show that the properties of scalar conditional expectations in fact 5.2.13 carry over to the matrix setting. Here's a partial list:

**Fact 5.2.14** If  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{Z}$  are random matrices and  $\mathbf{A}$  and  $\mathbf{B}$  are constant and conformable, then

- (i)  $\mathbb{E}[\mathbf{Y} | \mathbf{Z}]^\top = \mathbb{E}[\mathbf{Y}^\top | \mathbf{Z}]$ .
- (ii)  $\mathbb{E}[\mathbf{AX} + \mathbf{BY} | \mathbf{Z}] = \mathbf{A}\mathbb{E}[\mathbf{X} | \mathbf{Z}] + \mathbf{B}\mathbb{E}[\mathbf{Y} | \mathbf{Z}]$ .
- (iii)  $\mathbb{E}[\mathbb{E}[\mathbf{Y} | \mathbf{X}]] = \mathbb{E}[\mathbf{Y}]$  and  $\mathbb{E}[\mathbb{E}[\mathbf{Y} | \mathbf{X}, \mathbf{Z}] | \mathbf{X}] = \mathbb{E}[\mathbf{Y} | \mathbf{X}]$ .
- (iv) If  $\mathbf{X}$  and  $\mathbf{Y}$  are independent, then  $\mathbb{E}[\mathbf{Y} | \mathbf{X}] = \mathbb{E}[\mathbf{Y}]$ .
- (v) If  $g(\mathbf{X})$  is a matrix depending only on  $\mathbf{X}$ , then
  - (a)  $\mathbb{E}[g(\mathbf{X}) | \mathbf{X}] = g(\mathbf{X})$
  - (b)  $\mathbb{E}[g(\mathbf{X}) \mathbf{Y} | \mathbf{X}] = g(\mathbf{X})\mathbb{E}[\mathbf{Y} | \mathbf{X}]$  and  $\mathbb{E}[\mathbf{Y} g(\mathbf{X}) | \mathbf{X}] = \mathbb{E}[\mathbf{Y} | \mathbf{X}] g(\mathbf{X})$

Part (v) shows examples of conditional determinism in the matrix case.

### 5.3 Further Reading

Comprehensive treatments of multivariate distributions can be found in monographs such as Izenman (2008) and Joe (1997). Beautiful treatments of conditional expectation via orthogonal projection in  $L_2$  can be found in Williams (1991), Taylor (1997), Schilling (2005), and Çinlar (2011).

### 5.4 Exercises

**Ex. 5.4.1** Confirm the expression for the density  $q$  of  $y = \psi(x)$  in (4.32) in the case where  $\psi$  is strictly increasing.<sup>5</sup>

**Ex. 5.4.2** Let  $x_1$  and  $x_2$  be random variables with joint density  $p$  and marginals  $p_1$  and  $p_2$ . Show that  $x_1$  and  $x_2$  are independent in the sense that (5.19) holds whenever  $p(s_1, s_2) = p_1(s_1)p_2(s_2)$  for every  $(s_1, s_2) \in \mathbb{R}^2$ .<sup>6</sup>

**Ex. 5.4.3** Let  $p: \mathbb{R}^2 \rightarrow \mathbb{R}$  be defined by

$$p(s_1, s_2) = \begin{cases} \exp(s_1 + s_2) & \text{if } s_1 \leq 0 \text{ and } s_2 \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

Let  $x_1, x_2$  have joint density  $p$ . Show that  $x_1$  and  $x_2$  are independent.

**Ex. 5.4.4** Show that if  $\mathcal{L}(x) = N(0, 1)$ , then  $\mathcal{L}(x^2) = \chi^2(1)$ .<sup>7</sup>

**Ex. 5.4.5** Confirm fact 5.1.11: If  $\mathbf{x}$  and  $\mathbf{y}$  are independent, then  $\text{cov}[\mathbf{x}, \mathbf{y}] = 0$ .

**Ex. 5.4.6** Let  $x$  and  $y$  be independent uniform random variables on  $[0, 1]$ . Let  $z := \max\{x, y\}$ . Compute the CDF, density and mean of  $z$ .<sup>8</sup> In addition, compute the CDF of  $w := \min\{x, y\}$ .

**Ex. 5.4.7** Confirm fact 5.1.3 on page 127.

**Ex. 5.4.8** Let  $\mathbf{x}$  be a random vector. Confirm the claim in fact 5.1.2 that  $\text{var}[\mathbf{x}]$  is nonnegative definite.

**Ex. 5.4.9** Let  $\mathcal{L}(\mathbf{x}) = \mathcal{L}(x_1, \dots, x_N) = N(\mathbf{0}, \mathbf{I})$ .

---

5. Hint: Use fact 4.2.6 on page 111.

6. Hint: Apply (5.9) on page 132.

7. Hints: You need to match the density in example 4.2.12 when  $k = 1$ . Use the fact that  $\Gamma(1/2) = \sqrt{\pi}$ . Try to find the CDF of  $z^2$  and differentiate it. In finding the CDF, the solution to exercise 4.4.18 on page 117 might be helpful.

8. Hint: Fix  $s \in \mathbb{R}$  and compare the sets  $\{z \leq s\}$  and  $\{x \leq s\} \cap \{y \leq s\}$ . What is the relationship between these two sets?

- (i) What is the distribution of  $x_1^2/x_2^2$ ?
- (ii) What is the distribution of  $x_1[2/(x_2^2 + x_3^2)]^{1/2}$ ?
- (iii) What is the distribution of  $\|\mathbf{x}\|^2$ ?
- (iv) If  $\mathbf{a}$  is an  $N \times 1$  constant vector, what is the distribution of  $\mathbf{a}^\top \mathbf{x}$ ?

**Ex. 5.4.10** Let  $x$  and  $y$  be independent random variables taking values in  $\{0\} \cup \mathbb{N}$  with PMFs  $p = \{p(j)\}$  and  $q = \{q(j)\}$  respectively. (I'm using function rather than subscript style notation for this exercise to make the arguments a bit easier to follow.) The **convolution** of  $p$  and  $q$  is the PMF  $p \star q$  defined by

$$(p \star q)(j) = \sum_{k \geq 0} p(j-k)q(k) \quad (5.41)$$

Show that  $p \star q$  is the PMF of  $z := x + y$  when  $x$  and  $y$  are independent with  $\mathcal{L}(x) = p$  and  $\mathcal{L}(y) = q$ .

**Ex. 5.4.11** Prove fact 5.1.13 on page 137.

**Ex. 5.4.12** Fix  $z \in L_2$  with  $\mathbb{E}z = 0$  and let  $S := \{x \in L_2 : \text{cov}[x, z] = 0\}$ . Show that  $S$  is a linear subspace of  $L_2$ .

**Ex. 5.4.13** Let  $x \in L_2$  with  $\mathbb{E}x = \mu$  and  $\text{var}[x] = \sigma^2$ . Let  $S := \text{span}\{\mathbb{1}\}$ , as in example 5.2.5 on page 146. Let  $\mathbf{P} := \text{proj } S$  and let  $\mathbf{M}$  be the residual projection (see page 32) defined by  $\mathbf{M}x = x - \mathbf{P}x$ . Show that  $\|\mathbf{M}x\| = \sigma$ .

**Ex. 5.4.14** Let  $x, y \in L_2$ , let  $S := \text{span}\{x\}$ , and let  $\mathbf{P} := \text{proj } S$ . Show that

$$\mathbf{P}y = \mathbb{E}[x^2]^{-1} \mathbb{E}[xy] x \quad (5.42)$$

**Ex. 5.4.15** Prove the Cauchy–Schwarz inequality for random variables, which was first stated as fact 4.1.8 on page 97 as  $|\mathbb{E}[xy]| \leq (\mathbb{E}[x^2]\mathbb{E}[y^2])^{1/2}$  and repeated as part (iv) of fact 5.2.2. Use the results on orthogonal projections in  $L_2$  as found in §5.2.3.<sup>9</sup>

**Ex. 5.4.16** Prove that  $\hat{y}$  in (5.34) is the orthogonal projection of  $y$  onto  $\text{span}\{x\}$  under the conditions stated in fact 5.2.7.

**Ex. 5.4.17** Let  $\mathbf{x} \in \mathbb{R}^K$  be a random vector and let  $\|\cdot\|$  be the ordinary vector norm. Show that if  $\|\mathbf{x}\|$  is in  $L_2$ , then the random variable  $z = \ell(\mathbf{x})$  is in  $L_2$  whenever  $\ell: \mathbb{R}^K \rightarrow \mathbb{R}$  is linear.

---

9. Hint: Use (5.42) and review the solution to exercise 3.5.33.

**Ex. 5.4.18** Consider the setting of the population regression discussion in §5.2.3.1. In particular, let  $y$  be an element of  $L_2$ , let  $\mathbf{b}^*$  be the coefficients of the best linear predictor, as defined in (5.34), let  $S = \text{span}\{\mathbf{x}\}$ , and let  $f^*(\mathbf{x}) = \mathbb{E}[y | \mathbf{x}]$ . Show that  $\ell^*(\mathbf{x}) := \mathbf{x}^\top \mathbf{b}^*$  is the orthogonal projection of  $f^*(\mathbf{x})$  onto  $S$ .

**Ex. 5.4.19** Prove part (iii) of fact 5.2.13 when  $\mathcal{G} = \{\mathbf{x}\}$ .

**Ex. 5.4.20** Prove part (v) of fact 5.2.13.

**Ex. 5.4.21** Let  $\text{var}[y | x] := \mathbb{E}[y^2 | x] - (\mathbb{E}[y | x])^2$ . Show that

$$\text{var } y = \mathbb{E}[\text{var}[y | x]] + \text{var}[\mathbb{E}[y | x]]$$

**Ex. 5.4.22** Show that the conditional expectation of a constant  $\alpha$  is  $\alpha$ . In particular, using the results in fact 5.2.13 (page 151) as appropriate, show that if  $\alpha$  is a constant and  $\mathcal{G}$  is any information set, then  $\mathbb{E}[\alpha | \mathcal{G}] = \alpha$ .

**Ex. 5.4.23** Prove the claim in example 5.2.14.

### 5.4.1 Solutions to Selected Exercises

**Solution to Ex. 5.4.1.** Let  $x$ ,  $p$  and  $y = \psi(x)$  be as stated in fact 4.2.7 on page 111. Assume in addition that  $\psi$  is strictly increasing. Let  $F$  be the CDF of  $x$  and let  $G$  be the CDF of  $y$ . By fact 4.2.6 on page 111 we have  $G(s) = F(\psi^{-1}(s))$ . Differentiating gives

$$q(s) = G'(s) = p(\psi^{-1}(s)) \frac{d\psi^{-1}(s)}{ds}$$

Since  $\psi^{-1}$  is increasing, this agrees with (4.32).  $\square$

**Solution to Ex. 5.4.3.** Let  $p_1$  be the marginal density of  $x_1$  and  $p_2$  be the marginal density of  $x_2$ . It suffices to show that

$$p(s_1, s_2) = p_1(s_1)p_2(s_2) \quad \text{for all } (s_1, s_2) \in \mathbb{R}^2 \quad (5.43)$$

To see that this is the case, first we obtain the marginal distributions by integrating out the other variable. To start, let  $s_1 \leq 0$  and observe that

$$p_1(s_1) = \int_{-\infty}^0 p(s_1, s_2) ds_2 = \int_{-\infty}^0 e^{s_1} e^{s_2} ds_2 = e^{s_1} \int_{-\infty}^0 e^{s_2} ds_2 = e^{s_1}$$

Conversely, if  $s_1 > 0$ , then  $p(s_1, s_2) = 0$  for any  $s_2$ , and hence

$$p_1(s_1) = \int_{-\infty}^0 p(s_1, s_2) ds_2 = 0$$

From this and a symmetric argument in the case of  $p_2$ , we get

$$p_1(s_1) = \begin{cases} e^{s_1} & \text{if } s_1 \leq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad p_2(s_2) = \begin{cases} e^{s_2} & \text{if } s_2 \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

It follows that (5.43) is valid.  $\square$

**Solution to Ex. 5.4.4.** Let  $F(s)$  be the CDF of  $x^2$  and let  $f$  be the density. Let  $\Phi$  and  $\phi$  be the  $N(0,1)$  cdf and density respectively. Following arguments similar to the solution to exercise 4.4.18 we have

$$F(s) = \mathbb{P}\{x^2 \leq s\} = \mathbb{P}\{-\sqrt{s} \leq x \leq \sqrt{s}\} = 2\Phi(\sqrt{s}) - 1$$

Differentiating with respect to  $s$  gives  $f(s) = \phi(\sqrt{s})s^{-1/2}$ . Plugging in the definition of the  $N(0,1)$  density and using  $\Gamma(1/2) = \sqrt{\pi}$  gives the density in example 4.2.12 when  $k = 1$ .  $\square$

**Solution to Ex. 5.4.6.** As in the statement of the exercise,  $x$  and  $y$  are independent uniform random variables on  $[0,1]$ ,  $z := \max\{x, y\}$  and  $w := \min\{x, y\}$ . As a first step to the proofs, note that if  $a, b$  and  $c$  are three numbers, then

- $\max\{a, b\} \leq c$  if and only if  $a \leq c$  and  $b \leq c$ .
- $\min\{a, b\} \leq c$  if and only if  $a \leq c$  or  $b \leq c$ .

From these facts we can see that, for any  $s \in \mathbb{R}$ ,

$$\{z \leq s\} = \{x \leq s\} \cap \{y \leq s\} \quad \text{and} \quad \{w \leq s\} = \{x \leq s\} \cup \{y \leq s\}$$

Now, for  $s \in [0,1]$ , we have

$$\mathbb{P}\{z \leq s\} = \mathbb{P}[\{x \leq s\} \cap \{y \leq s\}] = \mathbb{P}\{x \leq s\}\mathbb{P}\{y \leq s\} = s^2$$

By differentiating, we get the density  $p(s) = 2s$ , and by integrating  $\int_0^1 sp(s) ds$ , we get  $\mathbb{E}[z] = 2/3$ . Finally, regarding the CDF of  $w$ , for  $s \in [0,1]$ , we have

$$\begin{aligned} \mathbb{P}\{w \leq s\} &= \mathbb{P}[\{x \leq s\} \cup \{y \leq s\}] \\ &= \mathbb{P}\{x \leq s\} + \mathbb{P}\{y \leq s\} - \mathbb{P}[\{x \leq s\} \cap \{y \leq s\}] \end{aligned}$$

Hence  $\mathbb{P}\{w \leq s\} = 2s - s^2$ .  $\square$

**Solution to Ex. 5.4.7.** Since  $\mathbf{z}^T \mathbf{A} \mathbf{z}$  has the form  $\sum_{i=1}^N \sum_{j=1}^N a_{ij} z_i z_j$ , under the stated as-

sumptions we have

$$\mathbb{E}[\mathbf{z}^\top \mathbf{A} \mathbf{z}] = \sum_{i=1}^N \sum_{j=1}^N a_{ij} \mathbb{E}[z_i z_j] = \text{trace } \mathbf{A} \quad \square$$

**Solution to Ex. 5.4.9.** The respective solutions are as follows:

- (i)  $\mathcal{L}(x_1^2/x_2^2) = F(1, 1)$  by fact 5.1.17.
- (ii)  $\mathcal{L}[x_1[2/(x_2^2 + x_3^2)]^{1/2}] = t(2)$  by fact 5.1.16.
- (iii)  $\mathcal{L}(\|\mathbf{x}\|^2) = \chi^2(N)$  by fact 5.1.18.
- (iv)  $\mathcal{L}(\mathbf{a}^\top \mathbf{x}) = \mathcal{N}(0, \|\mathbf{a}\|^2)$  by fact 5.1.5.

□

**Solution to Ex. 5.4.10.** Let  $p \star q$  be as defined in (5.41), let  $z := x + y$  where  $x$  and  $y$  are independent, and let  $\mathcal{L}(x) = p$  and  $\mathcal{L}(y) = q$ . The claim is that  $(p \star q)(j) = \mathbb{P}\{z = j\}$  for any  $j \geq 0$ . To see this, fix any such  $j$  and apply the law of total probability (page 88) to get

$$\mathbb{P}\{z = j\} = \sum_{k \geq 0} \mathbb{P}\{z = j \mid y = k\} \mathbb{P}\{y = k\} = \sum_{k \geq 0} \mathbb{P}\{x + k = j\} q(k)$$

Using  $\mathbb{P}\{x + k = j\} = \mathbb{P}\{x = j - k\} = p(j - k)$  completes the proof. □

**Solution to Ex. 5.4.11.** Let  $S$ ,  $\mathbf{P}$ ,  $\mathbf{M}$ , and  $\mathbf{z}$  be as in the statement of the question. In view of fact 5.1.12, to prove independence of  $\mathbf{P}\mathbf{z}$  and  $\mathbf{M}\mathbf{z}$  it suffices to show that  $\text{cov}[\mathbf{P}\mathbf{z}, \mathbf{M}\mathbf{z}] = \mathbf{0}$ . Since these vectors are zero-mean and  $\mathbf{M}$  is symmetric (see page 61), the covariance is equal to

$$\mathbb{E}[(\mathbf{P}\mathbf{z})(\mathbf{M}\mathbf{z})^\top] = \mathbb{E}[\mathbf{P}\mathbf{z}\mathbf{z}^\top \mathbf{M}] = \mathbf{P} \mathbb{E}[\mathbf{z}\mathbf{z}^\top] \mathbf{M} = \sigma^2 \mathbf{P}\mathbf{M} = \mathbf{0}$$

The final equality follows from fact 2.2.8 on page 33. □

**Solution to Ex. 5.4.12.** Let  $z$  and  $S$  be as described in the question.  $S$  is a linear subspace of  $L_2$  because if  $x, y \in S$  and  $\alpha, \beta \in \mathbb{R}$ , then

$$\text{cov}[\alpha x + \beta y, z] = \alpha \mathbb{E}[xz] + \beta \mathbb{E}[yz] = \alpha \text{cov}[x, z] + \beta \text{cov}[y, z] = 0 \quad \square$$

**Solution to Ex. 5.4.13.** We saw in example 5.2.5 that  $\mathbf{P}x = \mu \mathbb{1}$ , so  $\mathbf{M}x = x - \mu \mathbb{1}$ , or, written more conventionally,  $\mathbf{M}x = x - \mu$ . The norm of  $\mathbf{M}x$  is  $\|\mathbf{M}x\| = \sqrt{\mathbb{E}[(x - \mu)^2]} = \sigma$ , as was to be shown. □

**Solution to Ex. 5.4.14.** Let  $x, y, S$ , and  $\mathbf{P}$  be as in the statement of the exercise, and let  $\mathbf{P}y$  be as defined in (5.42). As a scalar multiple of  $x$  it's clear that  $\mathbf{P}y \in S$ , so we need only show that  $y - \mathbf{P}y \perp x$ , or, equivalently,  $\mathbb{E}[xy - x\mathbf{P}y] = 0$ . Substituting in the expression for  $\mathbf{P}y$  in (5.42) and using linearity of  $\mathbb{E}$  confirms this equality.  $\square$

**Solution to Ex. 5.4.16.** Assume the conditions of fact 5.2.7 and adopt the notation stated there. Let  $y$  be any element of  $L_2$ . In view of theorem 5.2.1, to show that  $\hat{y} = \mathbf{P}y$ , we need to show that  $\hat{y} \in S := \text{span}\{\mathbf{x}\}$  and that  $y - \hat{y} \perp S$ . The first part is trivial, since  $\hat{y} = \mathbf{x}^\top \mathbf{b}^*$  where  $\mathbf{b}^*$  is the vector  $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]^{-1}\mathbb{E}[\mathbf{x}y]$ . To prove the second part, it suffices to show that

$$\mathbb{E}[\mathbf{a}^\top \mathbf{x}y] = \mathbb{E}[\mathbf{a}^\top \mathbf{x}\mathbf{x}^\top \mathbf{b}^*]$$

Substituting in the definition of  $\mathbf{b}^*$  and some easy manipulations confirm that this equality holds.  $\square$

**Solution to Ex. 5.4.17.** Let  $z = \ell(\mathbf{x})$  where  $\ell$  is a linear function. The claim is that  $z$  is square integrable. By theorem 3.1.1 on page 48, we can write  $z$  as  $z = \mathbf{x}^\top \mathbf{a}$  for some  $\mathbf{a} \in \mathbb{R}^K$ . Applying the Cauchy–Schwarz inequality for vectors, we have  $z^2 \leq \|\mathbf{x}\|^2 \|\mathbf{a}\|^2$ . Since  $\|\mathbf{x}\|$  is in  $L_2$ , it follows that  $\mathbb{E}[z^2] < \infty$ .  $\square$

**Solution to Ex. 5.4.18.** Adopt the setting and notation of exercise 5.4.18. Let  $\mathcal{G}$  be the set of all  $\mathbf{x}$ -measurable functions in  $L_2$ . Let  $S = \text{span}\{\mathbf{x}\}$  = the linear  $\mathbf{x}$ -measurable functions in  $L_2$ . (As in §5.2.3.1, we are assuming that  $\|\mathbf{x}\|$  is in  $L_2$ , which assures us that elements of  $S$  are in fact in  $L_2$ . See ex. 5.4.17.) Clearly  $S \subset \mathcal{G}$ . The result now follows from fact 5.2.5 on page 146.  $\square$

**Solution to Ex. 5.4.19.** Let  $y$  be independent of  $\mathbf{x}$ . From the definition of conditional expectation, to show that  $\mathbb{E}[y | \mathbf{x}] = \mathbb{E}[y]$  we need to show that

- (i)  $\mathbb{E}[y]$  is  $\mathcal{G}$ -measurable, and
- (ii)  $\mathbb{E}[\mathbb{E}[y]g(\mathbf{x})] = \mathbb{E}[yg(\mathbf{x})]$  for any  $\mathcal{B}$ -measurable  $g$ .

Part (i) is trivial because  $\mathbb{E}[y]$  is constant (see example 5.2.11 on page 149). Regarding (ii), if  $g$  is any such function, then by facts 5.1.9 and 5.1.10 (see page 136) we have  $\mathbb{E}[yg(\mathbf{x})] = \mathbb{E}[y]\mathbb{E}[g(\mathbf{x})]$ . The result follows.  $\square$

**Solution to Ex. 5.4.20.** We need to show that if  $x$  is  $\mathcal{G}$ -measurable, then  $\mathbb{E}[xy | \mathcal{G}] = x\mathbb{E}[y | \mathcal{G}]$ . To confirm this, we must show that

- (i)  $x\mathbb{E}[y | \mathcal{G}]$  is  $\mathcal{G}$ -measurable, and
- (ii)  $\mathbb{E}[x\mathbb{E}[y | \mathcal{G}]z] = \mathbb{E}[xyz]$  for any  $z \in L_2(\mathcal{G})$ .

Regarding (i),  $\mathbb{E}[y | \mathcal{G}]$  is  $\mathcal{G}$ -measurable by definition, and  $x$  is  $\mathcal{G}$ -measurable by assumption, so  $x\mathbb{E}[y | \mathcal{G}]$  is  $\mathcal{G}$ -measurable by fact 5.2.8 on page 149. Regarding (ii), fix  $z \in L_2(\mathcal{G})$ , and let  $u := xz$ . Since  $x \in L_2(\mathcal{G})$ , we have  $u \in L_2(\mathcal{G})$ . We need to show that

$$\mathbb{E}[\mathbb{E}[y | \mathcal{G}]u] = \mathbb{E}[yu]$$

Since  $u \in L_2(\mathcal{G})$ , this is immediate from the definition of  $\mathbb{E}[y | \mathcal{G}]$ .  $\square$

**Solution to Ex. 5.4.22.** By fact 5.2.13 (page 151), we know that if  $\alpha$  is  $\mathcal{G}$ -measurable, then  $\mathbb{E}[\alpha | \mathcal{G}] = \alpha$ . Example 5.2.11 on page 149 tells us that  $\alpha$  is indeed  $\mathcal{G}$ -measurable.  $\square$

**Solution to Ex. 5.4.23.** As in example 5.2.14, let  $x$  and  $y$  be random variables where  $p(y | x)$  is the conditional density of  $y$  given  $x$ . Let  $g(x) := \int tp(t | x) dt$ . The claim is that  $\mathbb{E}[y | x] = g(x)$ . To prove this, we need to show that  $g(x)$  is  $x$ -measurable, and that

$$\mathbb{E}[g(x)h(x)] = \mathbb{E}[yh(x)] \quad \text{for any } \mathcal{B}\text{-measurable } h \quad (5.44)$$

The first claim is obvious. Regarding (5.44), let  $h$  be any such function. Using the notation in (5.26) on page 142, we can write

$$\begin{aligned} \mathbb{E}[g(x)h(x)] &= \mathbb{E} \left[ \int tp(t | x) dt h(x) \right] \\ &= \int \int tp(t | s) dt h(s) p(s) ds \\ &= \int \int t \frac{p(s, t)}{p(s)} dt h(s) p(s) ds = \int \int t h(s) p(s, t) dt ds \end{aligned}$$

This is equal to the right-hand side of (5.44), and the proof is done.  $\square$