for scalar λ, which can be rewritten

$$(\mathbf{M} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}, \tag{30}$$

where $\mathbf{I}$ the identity matrix and $\mathbf{0}$ is the zero vector. The solution vector $\mathbf{x} = \mathbf{e}_i$ and corresponding scalar $\lambda = \lambda_i$ are called the *eigenvector* and associated *eigenvalue*, respectively. If $\mathbf{M}$ is real and symmetric, there are d (possibly nondistinct) solution vectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d\}$, each with an associated eigenvalue $\{\lambda_1, \lambda_2, \ldots, \lambda_d\}$. Under multiplication by $\mathbf{M}$ the eigenvectors are changed only in magnitude, not direction:

$$\mathbf{M}\mathbf{e}_j = \lambda_j \mathbf{e}_j. \tag{31}$$

If $\mathbf{M}$ is diagonal, then the eigenvectors are parallel to the coordinate axes.

CHARACTERISTIC EQUATION
SECULAR EQUATION

One method of finding the eigenvectors and eigenvalues is to solve the *characteristic equation* (or *secular equation*,)

$$|\mathbf{M} - \lambda\mathbf{I}| = \lambda^d + a_1\lambda^{d-1} + \cdots + a_{d-1}\lambda + a_d = 0, \tag{32}$$

for each of its d (possibly nondistinct) roots λ_j. For each such root, we then solve a set of linear equations to find its associated eigenvector $\mathbf{e}_j$.

Finally, it can be shown that the trace of a matrix is just the sum of the eigenvalues and the determinant of a matrix is just the product of its eigenvalues:

$$\mathrm{tr}[\mathbf{M}] = \sum_{i=1}^{d} \lambda_i \quad \text{and} \quad |\mathbf{M}| = \prod_{i=1}^{d} \lambda_i. \tag{33}$$

If a matrix is diagonal, then its eigenvalues are simply the nonzero entries on the diagonal, and the eigenvectors are the unit vectors parallel to the coordinate axes.

A.3 LAGRANGE OPTIMIZATION

Suppose we seek the position $\mathbf{x}_0$ of an extremum of a scalar-valued function $f(\mathbf{x})$, subject to some constraint. If a constraint can be expressed in the form $g(\mathbf{x}) = 0$, then we can find the extremum of $f(\mathbf{x})$ as follows. First we form the Lagrangian function

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \underbrace{\lambda g(\mathbf{x})}_{=0}, \tag{34}$$

UNDETERMINED MULTIPLIER

where λ is a scalar called the Lagrange *undetermined multiplier*. We convert this constrained optimization problem into an unconstrained problem by taking the derivative,

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \lambda \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = 0, \tag{35}$$

and using standard methods from calculus to solve the resulting equations for λ and the extremizing value of $\mathbf{x}$. (Note that the term $\lambda \partial g/\partial \mathbf{x}$ does not vanish, in general.) The solution gives the $\mathbf{x}$ position of the extremum, and it is a simple matter of substitution to find the extreme value of $f(\cdot)$ under the constraints.

A.4 PROBABILITY THEORY

A.4.1 Discrete Random Variables

Let x be a discrete random variable that can assume any of the finite number m of different values in the set $\mathcal{X} = \{v_1, v_2, \ldots, v_m\}$. We denote by p_i the probability that x assumes the value v_i:

$$p_i = \Pr[x = v_i], \qquad i = 1, \ldots, m. \tag{36}$$

Then the probabilities p_i must satisfy the following two conditions:

$$p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} p_i = 1. \tag{37}$$

PROBABILITY MASS FUNCTION

Sometimes it is more convenient to express the set of probabilities $\{p_1, p_2, \ldots, p_m\}$ in terms of the *probability mass function* $P(x)$, which must satisfy the following conditions:

$$P(x) \geq 0, \quad \text{and} \quad \sum_{x \in \mathcal{X}} P(x) = 1. \tag{38}$$

A.4.2 Expected Values

MEAN

The *expected value*, *mean*, or *average* of the random variable x is defined by

$$\mathcal{E}[x] = \mu = \sum_{x \in \mathcal{X}} x P(x) = \sum_{i=1}^{m} v_i p_i. \tag{39}$$

If one thinks of the probability mass function as defining a set of point masses, with p_i being the mass concentrated at $x = v_i$, then the expected value μ is just the center of mass. Alternatively, we can interpret μ as the arithmetic average of the values in a large random sample. More generally, if $f(x)$ is any function of x, the expected value of f is defined by

$$\mathcal{E}[f(x)] = \sum_{x \in \mathcal{X}} f(x) P(x). \tag{40}$$

Note that the process of forming an expected value is *linear*, in that if α_1 and α_2 are arbitrary constants, then we have

$$\mathcal{E}[\alpha_1 f_1(x) + \alpha_2 f_2(x)] = \alpha_1 \mathcal{E}[f_1(x)] + \alpha_2 \mathcal{E}[f_2(x)]. \tag{41}$$

EXPECTATION OPERATOR
SECOND MOMENT
VARIANCE

It is sometimes convenient to think of $\mathcal{E}$ as an operator—the (linear) *expectation operator*. Two important special-case expectations are the *second moment* and the *variance*:

$$\mathcal{E}[x^2] = \sum_{x \in \mathcal{X}} x^2 P(x) \tag{42}$$

$$\mathrm{Var}[x] = \sigma^2 = \mathcal{E}[(x - \mu)^2] = \sum_{x \in \mathcal{X}} (x - \mu)^2 P(x), \tag{43}$$

STANDARD DEVIATION

where σ is the *standard deviation* of x. The variance can be viewed as the moment of inertia of the probability mass function. The variance is never negative, and it is zero if and only if all of the probability mass is concentrated at one point.

The standard deviation is a simple but valuable measure of how far values of x are likely to depart from the mean. Its very name suggests that it is the standard or typical amount one should expect a randomly drawn value for x to deviate or differ from μ. *Chebyshev's inequality* (or the Bienaymé-Chebyshev inequality) provides a mathematical relation between the standard deviation and $|x - \mu|$:

CHEBYSHEV'S INEQUALITY

$$\Pr[|x - \mu| > n\sigma] \leq \frac{1}{n^2}. \tag{44}$$

This inequality is not a tight bound (and it is useless for $n < 1$); a more practical rule of thumb, which strictly speaking is true only for the normal distribution, is that 68% of the values will lie within one, 95% within two, and 99.7% within three standard deviations of the mean (cf. Fig. A.1, ahead). Nevertheless, Chebyshev's inequality shows the strong link between the standard deviation and the spread of a distribution. In addition, it suggests that $|x - \mu|/\sigma$ is a meaningful normalized measure of the distance from x to the mean (cf. Section A.4.12).

By expanding the quadratic in Eq. 43, it is easy to prove the useful formula

$$\mathrm{Var}[x] = \mathcal{E}[x^2] - (\mathcal{E}[x])^2. \tag{45}$$

Note that, unlike the mean, the variance is *not* linear. In particular, if $y = \alpha x$, where α is a constant, then $\mathrm{Var}[y] = \alpha^2 \mathrm{Var}[x]$. Moreover, the variance of the sum of two random variables is usually *not* the sum of their variances. However, as we shall see below, variances do add when the variables involved are statistically independent.

In the simple but important special case in which x is binary-valued (say, $v_1 = 0$ and $v_2 = 1$), we can obtain simple formulas for μ and σ. If we let $p = \Pr[x = 1]$, then it is easy to show that

$$\mu = p \qquad \text{and} \qquad \sigma = \sqrt{p(1-p)}. \tag{46}$$

A.4.3 Pairs of Discrete Random Variables

PRODUCT SPACE

Let x and y be random variables which can take on values in $\mathcal{X} = \{v_1, v_2, \ldots, v_m\}$, and $\mathcal{Y} = \{w_1, w_2, \ldots, w_n\}$, respectively. We can think of (x, y) as a vector or a point in the *product space* of x and y. For each possible pair of values (v_i, w_j) we have a *joint probability* $p_{ij} = \Pr[x = v_i, y = w_j]$. These mn joint probabilities p_{ij} are nonnegative and sum to 1. Alternatively, we can define a *joint probability mass function* $P(x, y)$ for which

$$P(x, y) \geq 0 \qquad \text{and} \qquad \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1. \tag{47}$$

MARGINAL DISTRIBUTION

The joint probability mass function is a complete characterization of the pair of random variables (x, y); that is, everything we can compute about x and y, individually or together, can be computed from $P(x, y)$. In particular, we can obtain the separate *marginal distributions* for x and y by summing over the unwanted variable:

$$\begin{aligned} P_x(x) &= \sum_{y \in \mathcal{Y}} P(x, y) \\ P_y(y) &= \sum_{x \in \mathcal{X}} P(x, y). \end{aligned} \tag{48}$$

We will occasionally use subscripts, as in Eq. 48, to emphasize the fact that $P_x(x)$ has a different functional form than $P_y(y)$. It is common to omit them and write simply $P(x)$ and $P(y)$ whenever the context makes it clear that these are in fact two different functions—rather than the same function merely evaluated with different values for the argument.

A.4.4 Statistical Independence

Variables x and y are said to be *statistically independent* if and only if

$$P(x, y) = P_x(x) P_y(y). \tag{49}$$

We can understand such independence as follows. Suppose that $p_i = \Pr[x = v_i]$ is the fraction of the time that $x = v_i$, and $q_j = \Pr[y = w_j]$ is the fraction of the time that $y = w_j$. Consider those situations where $x = v_i$. If it is still true that the fraction of those situations in which $y = w_j$ is the same value q_j, it follows that knowing the value of x did not give us any additional knowledge about the possible values of y; in that sense y is independent of x. Finally, if x and y are statistically independent, it is clear that the fraction of the time that the specific pair of values (v_i, w_j) occurs must be the product of the fractions $p_i q_j = P(v_i)P(w_j)$ as we shall explore in Section A.4.6.

A.4.5 Expected Values of Functions of Two Variables

In the natural extension of Section A.4.2, we define the expected value of a function $f(x, y)$ of two random variables x and y by

$$\mathcal{E}[f(x, y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x, y) P(x, y), \tag{50}$$

and as before the expectation operator $\mathcal{E}$ is linear:

$$\mathcal{E}[\alpha_1 f_1(x, y) + \alpha_2 f_2(x, y)] = \alpha_1 \mathcal{E}[f_1(x, y)] + \alpha_2 \mathcal{E}[f_2(x, y)]. \tag{51}$$

The means (first moments) and variances (second moments) are

$$\begin{aligned} \mu_x &= \mathcal{E}[x] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x P(x, y) \\ \mu_y &= \mathcal{E}[y] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y P(x, y) \\ \sigma_x^2 &= \mathrm{Var}[x] = \mathcal{E}[(x - \mu_x)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)^2 P(x, y) \\ \sigma_y^2 &= \mathrm{Var}[y] = \mathcal{E}[(y - \mu_y)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (y - \mu_y)^2 P(x, y). \end{aligned} \tag{52}$$

COVARIANCE

An important new "cross-moment" can now be defined, the *covariance* of x and y:

$$\sigma_{xy} = \mathcal{E}[(x - \mu_x)(y - \mu_y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)(y - \mu_y) P(x, y). \tag{53}$$

Using vector notation, we can summarize Eqs. 52 and 53 as

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \sum_{\mathbf{x}\in\{\mathcal{X}\mathcal{Y}\}} \mathbf{x}P(\mathbf{x}) \tag{54}$$

$$\boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t], \tag{55}$$

where $\{\mathcal{X}\mathcal{Y}\}$ respresents the space of all possible values for all components of $\mathbf{x}$ and $\boldsymbol{\Sigma}$ is the covariance matrix (cf., Section A.4.9).

UNCORRELATED

The covariance is one measure of the degree of statistical dependence between x and y. If x and y are statistically independent, then $\sigma_{xy} = 0$. If $\sigma_{xy} = 0$, the variables x and y are said to be *uncorrelated*. It does *not* follow that uncorrelated variables must be statistically independent—covariance is just one measure of dependence. However, it is a fact that uncorrelated variables are statistically independent if they have a multivariate normal distribution, and in practice statisticians often treat uncorrelated variables as if they were statistically independent. If α is a constant and $y = \alpha x$, which is a case of strong statistical dependence, it is also easy to show that $\sigma_{xy} = \alpha\sigma_x^2$. Thus, the covariance is positive if x and y both increase or decrease together, and is negative if y decreases when x increases.

CAUCHY-SCHWARZ INEQUALITY

There is an important *Cauchy-Schwarz inequality* for the variances σ_x and σ_y and the covariance σ_{xy}. It can be derived by observing that the variance of a random variable is never negative, and thus the variance of $\lambda x + y$ must be nonnegative no matter what the value of the scalar λ. This leads to the famous inequality

$$\sigma_{xy}^2 \le \sigma_x^2\sigma_y^2, \tag{56}$$

which is analogous to the vector inequality $(\mathbf{x}^t\mathbf{y})^2 \le \|\mathbf{x}\|^2\,\|\mathbf{y}\|^2$ given in Eq. 8.

CORRELATION COEFFICIENT

The *correlation coefficient*, defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}, \tag{57}$$

is a normalized covariance, and must always be between -1 and $+1$. If $\rho = +1$, then x and y are maximally positively correlated, while if $\rho = -1$, they are maximally negatively correlated. If $\rho = 0$, the variables are uncorrelated. It is common practice to consider variables to be uncorrelated for practical purposes if the magnitude of their correlation coefficient is below some threshold, such as 0.05, although the threshold that makes sense does depend on the actual situation.

If x and y are statistically independent, then for any two functions f and g we obtain

$$\mathcal{E}[f(x)g(y)] = \mathcal{E}[f(x)]\mathcal{E}[g(y)], \tag{58}$$

a result which follows from the definition of statistical independence and expectation. Note that if $f(x) = x - \mu_x$ and $g(y) = y - \mu_y$, this theorem again shows that $\sigma_{xy} = \mathcal{E}[(x-\mu_x)(y-\mu_y)]$ is zero if x and y are statistically independent.

A.4.6 Conditional Probability

When two variables are statistically dependent, knowing the value of one of them lets us get a better estimate of the value of the other one. This is expressed by the

following definition of the *conditional probability* of x given y:

$$\Pr[x = v_i | y = w_j] = \frac{\Pr[x = v_i, y = w_j]}{\Pr[y = w_j]}, \tag{59}$$

or, in terms of mass functions,

$$P(x|y) = \frac{P(x, y)}{P(y)}. \tag{60}$$

Note that if x and y are statistically independent, this gives $P(x|y) = P(x)$. That is, when x and y are independent, knowing the value of y gives you no information about x that you didn't already know from its marginal distribution $P(x)$.

Consider a simple illustration of a two-variable binary case where both x and y are either 0 or 1. Suppose that a large number n of pairs of xy-values are randomly produced. Let n_{ij} be the number of pairs in which we find $x = i$ and $y = j$, that is, we see the (0, 0) pair n_{00} times, the (0, 1) pair n_{01} times, and so on, where $n_{00} + n_{01} + n_{10} + n_{11} = n$. Suppose we pull out those pairs where $y = 1$—that is, the (0, 1) pairs and the (1, 1) pairs. Clearly, the fraction of those cases in which x is also 1 is

$$\frac{n_{11}}{n_{01}+n_{11}} = \frac{n_{11}/n}{(n_{01}+n_{11})/n}. \tag{61}$$

Intuitively, this is what we would like to get for $P(x|y)$ when $y = 1$ and n is large. And, indeed, this is what we do get, because n_{11}/n is approximately $P(x, y)$ and $\frac{n_{11}/n}{(n_{01}+n_{11})/n}$ is approximately $P(y)$ for large n.

A.4.7 The Law of Total Probability and Bayes Rule

The *Law of Total Probability* states that if an event A can occur in m different ways $A_1, A_2, \ldots, A_m$ and if these m subevents are *mutually exclusive*—that is, cannot occur at the same time—then the probability of A occurring is the sum of the probabilities of the subevents A_i. In particular, the random variable y can assume the value y in m different ways—with $x = v_1$, $x = v_2$, ..., and $x = v_m$. Because these possibilities are mutually exclusive, it follows from the Law of Total Probability that $P(y)$ is the sum of the joint probability $P(x, y)$ over all possible values for x. Formally we have

$$P(y) = \sum_{x\in\mathcal{X}} P(x, y). \tag{62}$$

But from the definition of the conditional probability $P(y|x)$ we have

$$P(x, y) = P(y|x)P(x), \tag{63}$$

and after rewriting Eq. 63 with x and y exchanged and some simple algebra, we obtain

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x\in\mathcal{X}} P(y|x)P(x)}, \tag{64}$$

or in words we have

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}},$$

where these terms are discussed more fully in Chapter 2.

Equation 64 is called *Bayes rule*. Note that the denominator, which is just $P(y)$, is obtained by summing the numerator over all x values. By writing the denominator in this form we emphasize the fact that everything on the right-hand side of the equation is conditioned on x. If we think of x as the important variable, then we can say that the shape of the distribution $P(x|y)$ depends only on the numerator $P(y|x)P(x)$; the denominator is just a normalizing factor, sometimes called the *evidence*, needed to ensure that the $P(x|y)$ sum to one.

EVIDENCE

The standard interpretation of Bayes rule is that it "inverts" statistical connections, turning $P(y|x)$ into $P(x|y)$. Suppose that we think of x as a "cause" and y as an "effect" of that cause. That is, we assume that if the cause x is present, it is easy to determine the probability of the effect y being observed; the conditional probability function $P(y|x)$—the *likelihood*—specifies this probability explicitly. If we observe the effect y, it might not be so easy to determine the cause x, because there might be several different causes, each of which could produce the same observed effect. However, Bayes rule makes it easy to determine $P(x|y)$, provided that we know both $P(y|x)$ and the so-called *prior probability* $P(x)$, the probability of x before we make any observations about y. Said slightly differently, Bayes rule shows how the probability distribution for x changes from the *prior distribution* $P(x)$ before anything is observed about y to the *posterior distribution* $P(x|y)$ once we have observed the value of y.

LIKELIHOOD

PRIOR

POSTERIOR DISTRIBUTION

A.4.8 Vector Random Variables

To extend these results from two variables x and y to d variables $x_1, x_2, \ldots, x_d$, it is convenient to employ vector notation. As given by Eq. 47, the joint probability mass function $P(\mathbf{x})$ satisfies $P(\mathbf{x}) \geq 0$ and $\sum P(\mathbf{x}) = 1$, where the sum extends over all possible values for the vector $\mathbf{x}$. Note that $P(\mathbf{x})$ is a function of d variables and can be a very complicated, multidimensional function. However, if the random variables x_i are statistically independent, it reduces to the product

$$\begin{aligned} P(\mathbf{x}) &= P_{x_1}(x_1)P_{x_2}(x_2)\cdots P_{x_d}(x_d) \\ &= \prod_{i=1}^{d} P_{x_i}(x_i), \end{aligned} \tag{65}$$

where we have used the subscripts just to emphasize the fact that the marginal distributions will generally have a different form. Here the separate marginal distributions $P_{x_i}(x_i)$ can be obtained by summing the joint distribution over the other variables. In addition to these univariate marginals, other marginal distributions can be obtained by this use of the Law of Total Probability. For example, suppose we have $P(x_1, x_2, x_3, x_4, x_5)$ and we want $P(x_1, x_4)$; we merely calculate

$$P(x_1, x_4) = \sum_{x_2}\sum_{x_3}\sum_{x_5} P(x_1, x_2, x_3, x_4, x_5). \tag{66}$$

One can define many different conditional distributions, such as $P(x_1, x_2|x_3)$ or $P(x_2|x_1, x_4, x_5)$. For example,

$$P(x_1, x_2|x_3) = \frac{P(x_1, x_2, x_3)}{P(x_3)}, \tag{67}$$

where all of the joint distributions can be obtained from $P(\mathbf{x})$ by summing out the unwanted variables. If instead of scalars we have vector variables, then these conditional distributions can also be written as

$$P(\mathbf{x}_1|\mathbf{x}_2) = \frac{P(\mathbf{x}_1, \mathbf{x}_2)}{P(\mathbf{x}_2)}, \tag{68}$$

and likewise, in vector form, Bayes rule becomes

$$P(\mathbf{x}_1|\mathbf{x}_2) = \frac{P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_1)}{\sum_{\mathbf{x}_1} P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_1)}. \tag{69}$$

A.4.9 Expectations, Mean Vectors and Covariance Matrices

The expected value of a vector is defined to be the vector whose components are the expected values of the original components. Thus, if $\mathbf{f}(\mathbf{x})$ is an n-dimensional, vector-valued function of the d-dimensional random vector $\mathbf{x}$,

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{bmatrix}, \tag{70}$$

then the expected value of $\mathbf{f}$ is defined by

$$\mathcal{E}[\mathbf{f}] = \begin{bmatrix} \mathcal{E}[f_1(\mathbf{x})] \\ \mathcal{E}[f_2(\mathbf{x})] \\ \vdots \\ \mathcal{E}[f_n(\mathbf{x})] \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{f}(\mathbf{x})P(\mathbf{x}). \tag{71}$$

MEAN VECTOR

In particular, the d-dimensional *mean vector* $\boldsymbol{\mu}$ is defined by

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \begin{bmatrix} \mathcal{E}[x_1] \\ \mathcal{E}[x_2] \\ \vdots \\ \mathcal{E}[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{x}P(\mathbf{x}). \tag{72}$$

COVARIANCE MATRIX

Similarly, the *covariance matrix* $\boldsymbol{\Sigma}$ is defined as the (square) matrix whose ijth element σ_{ij} is the covariance of x_i and x_j:

$$\sigma_{ij} = \sigma_{ji} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)] \qquad i, j = 1 \ldots d, \tag{73}$$

as we saw in the two-variable case of Eq. 53. Therefore, in expanded form we have

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathcal{E}[(x_1-\mu_1)(x_1-\mu_1)] & \mathcal{E}[(x_1-\mu_1)(x_2-\mu_2)] & \dots & \mathcal{E}[(x_1-\mu_1)(x_d-\mu_d)] \\ \mathcal{E}[(x_2-\mu_2)(x_1-\mu_1)] & \mathcal{E}[(x_2-\mu_2)(x_2-\mu_2)] & \dots & \mathcal{E}[(x_2-\mu_2)(x_d-\mu_d)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}[(x_d-\mu_d)(x_1-\mu_1)] & \mathcal{E}[(x_d-\mu_d)(x_2-\mu_2)] & \dots & \mathcal{E}[(x_d-\mu_d)(x_d-\mu_d)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d^2 \end{bmatrix}. \tag{74}$$

We can use the vector product $(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t$ to write the covariance matrix as

$$\boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t]. \tag{75}$$

Thus, $\boldsymbol{\Sigma}$ is symmetric, and its diagonal elements are just the variances of the individual elements of $\mathbf{x}$, which can never be negative; the off-diagonal elements are the covariances, which can be positive or negative. If the variables are statistically independent, the covariances are zero, and the covariance matrix is diagonal. The analog to the Cauchy-Schwarz inequality comes from recognizing that if $\mathbf{w}$ is any d-dimensional vector, then the variance of $\mathbf{w}^t\mathbf{x}$ can never be negative. This leads to the requirement that the quadratic form $\mathbf{w}^t\boldsymbol{\Sigma}\mathbf{w}$ never be negative. Matrices for which this is true are said to be *positive semidefinite*; thus, the covariance matrix $\boldsymbol{\Sigma}$ must be positive semidefinite. It can be shown that this is equivalent to the requirement that none of the eigenvalues of $\boldsymbol{\Sigma}$ can be negative.

A.4.10 Continuous Random Variables

PROBABILITY DENSITY

When the random variable x can take values in the continuum, it no longer makes sense to talk about the probability that x has a particular value, such as 2.5136, because the probability of any particular exact value will almost always be zero. Rather, we talk about the probability that x falls in some interval (a, b); instead of having a probability mass function $P(x)$, we have a *probability density function* $p(x)$. The density has the property that

$$\Pr[x \in (a,b)] = \int_a^b p(x)\,dx. \tag{76}$$

The name *density* comes by analogy with material density. If we consider a small interval $(a, a+\Delta x)$ over which $p(x)$ is essentially constant, having value $p(a)$, we see that $p(a) = \Pr[x \in (a, a+\Delta x)]/\Delta x$. That is, the probability density at $x = a$ is the probability mass $\Pr[x \in (a, a+\Delta x)]$ per unit distance. It follows that the probability density function must satisfy

$$p(x) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} p(x)\,dx = 1. \tag{77}$$

In general, most of the definitions and formulas for discrete random variables carry over to continuous random variables with sums replaced by integrals. In particular, the expected value, mean, and variance for a continuous random variable are defined by

$$\mathcal{E}[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)\,dx$$

$$\mu = \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x)\,dx \tag{78}$$

$$\mathrm{Var}[x] = \sigma^2 = \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,dx,$$

and, as in Eq. 45, the variance obeys $\sigma^2 = \mathcal{E}[x^2] - (\mathcal{E}[x])^2$.

The multivariate situation is similarly handled with continuous random vectors $\mathbf{x}$. The probability density function $p(\mathbf{x})$ must satisfy

$$p(\mathbf{x}) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} p(\mathbf{x})\,d\mathbf{x} = 1, \tag{79}$$

where the integral is understood to be a d-fold, multiple integral and where $d\mathbf{x}$ is the element of d-dimensional volume $d\mathbf{x} = dx_1 dx_2 \cdots dx_d$. The corresponding moments for a general n-dimensional vector-valued function are

$$\mathcal{E}[\mathbf{f}(\mathbf{x})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x})p(\mathbf{x})\,dx_1 dx_2 \cdots dx_d = \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x})p(\mathbf{x})\,d\mathbf{x} \tag{80}$$

and for the particular d-dimensional functions as above, we have

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x}p(\mathbf{x})\,d\mathbf{x} \tag{81}$$

$$\boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int_{-\infty}^{\infty} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})\,d\mathbf{x}.$$

If the components of $\mathbf{x}$ are statistically independent, then the joint probability density function factors as

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_{x_i}(x_i) \tag{82}$$

and the covariance matrix is diagonal.

Conditional probability density functions are defined just as conditional mass functions. Thus, for example, the density for x given y is given by

$$p(x|y) = \frac{p(x, y)}{p(y)} \tag{83}$$

and Bayes rule for density functions is

$$p(x|y) = \frac{p(y|x)p(x)}{\int_{-\infty}^{\infty} p(y|x)p(x)\,dx}, \tag{84}$$

and likewise for the vector case.

Occasionally we will need to take the expectation with respect to a subset of the variables, and in that case we must show this as a subscript—for instance,

$$\mathcal{E}_{x_1}[f(x_1, x_2)] = \int_{-\infty}^{\infty} f(x_1, x_2)p(x_1)\,dx_1. \tag{85}$$

A.4.11 Distributions of Sums of Independent Random Variables

It frequently happens that we know the densities for two independent random variables x and y, and we need to know the density of their sum $z = x + y$. It is easy to obtain the mean and the variance of this sum:

$$\begin{aligned}
\mu_z &= \mathcal{E}[z] = \mathcal{E}[x + y] = \mathcal{E}[x] + \mathcal{E}[y] = \mu_x + \mu_y, \\
\sigma_z^2 &= \mathcal{E}[(z - \mu_z)^2] = \mathcal{E}[(x + y - (\mu_x + \mu_y))^2] = \mathcal{E}[((x - \mu_x) + (y - \mu_y))^2] \\
&= \mathcal{E}[(x - \mu_x)^2] + 2\underbrace{\mathcal{E}[(x - \mu_x)(y - \mu_y)]}_{=0} + \mathcal{E}[(y - \mu_y)^2] \\
&= \sigma_x^2 + \sigma_y^2,
\end{aligned} \tag{86}$$

where we have used the fact that the cross-term factors into $\mathcal{E}[x - \mu_x]\mathcal{E}[y - \mu_y]$ when x and y are independent; in this case the product is manifestly zero, because each of the component expectations vanishes. Thus, the mean of the sum of two independent random variables is the sum of their means, and the variance of their sum is the sum of their variances. If the variables are random *yet not independent*—for instance $y = -x$, where x is a random variable—then the variance is not the sum of the component variances.

It is only slightly more difficult to work out the exact probability density function for $z = x + y$ from the separate density functions for x and y. The probability that z is between ζ and $\zeta + \Delta z$ can be found by integrating the joint density $p(x, y) = p_x(x)p_y(y)$ over the thin strip in the xy-plane between the lines $x + y = \zeta$ and $x + y = \zeta + \Delta z$. It follows that, for small Δz,

$$\Pr[\zeta < z < \zeta + \Delta z] = \left[\int_{-\infty}^{\infty} p(x)p(\zeta - x)\,dx\right]\Delta z, \tag{87}$$

CONVOLUTION

and hence that the probability density function for the sum is the *convolution* of the probability density functions for the components:

$$p(z) = p_x(x) \star p_y(y) = \int_{-\infty}^{\infty} p_x(x)p_y(z - x)\,dx. \tag{88}$$

As one would expect, these results generalize. It is not hard to show that:

- The mean of the sum of d independent random variables $x_1, x_2, \ldots, x_d$ is the sum of their means. (In fact the variables need not be independent for this to hold.)
- The variance of the sum is the sum of their variances.
- The probability density function for the sum is the convolution of the separate density functions:

$$p(z) = p(x_1) \star p(x_2) \star \cdots \star p(x_d). \tag{89}$$

A.4.12 Normal Distributions

CENTRAL LIMIT THEOREM

GAUSSIAN

One of the most important results of probability theory is the *Central Limit Theorem*, which states that, under various conditions, the distribution for the sum of d independent random variables approaches a particular limiting form known as the *normal distribution*. As such, the *normal* or *Gaussian* probability density function is very important, both for theoretical and practical reasons. In one dimension, it is defined by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2((x-\mu)^2/\sigma^2)}. \tag{90}$$

The normal density is traditionally described as a "bell-shaped curve"; it is completely determined by the numerical values for two parameters, the mean μ and the variance σ^2. This is often emphasized by writing $p(x) \sim N(\mu, \sigma^2)$, which is read as "$x$ is distributed normally with mean μ and variance σ^2." The distribution is symmetrical about the mean, the peak occurring at $x = \mu$ and the width of the "bell" is proportional to the standard deviation σ. The parameters of a normal density in Eq. 90 satisfy the following equations:

$$\begin{aligned}
\mathcal{E}[1] &= \int_{-\infty}^{\infty} p(x)\,dx = 1 \\
\mathcal{E}[x] &= \int_{-\infty}^{\infty} x\,p(x)\,dx = \mu \\
\mathcal{E}[(x - \mu)^2] &= \int_{-\infty}^{\infty} (x - \mu)^2 p(x)\,dx = \sigma^2.
\end{aligned} \tag{91}$$

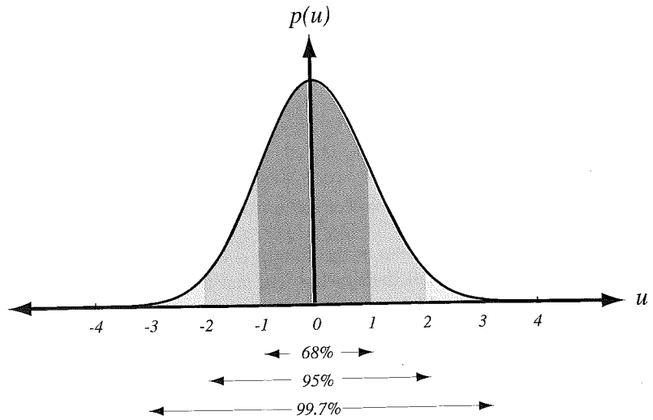

FIGURE A.1. A one-dimensional Gaussian distribution, $p(u) \sim N(0, 1)$, has 68% of its probability mass in the range $|u| \leq 1$, 95% in the range $|u| \leq 2$, and 99.7% in the range $|u| \leq 3$.

Normally distributed data points tend to cluster about the mean. Numerically, the probabilities obey

$$\begin{aligned} \Pr[|x-\mu| \leq \sigma] &\simeq 0.68 \\ \Pr[|x-\mu| \leq 2\sigma] &\simeq 0.95 \\ \Pr[|x-\mu| \leq 3\sigma] &\simeq 0.997, \end{aligned} \tag{92}$$

as shown in Fig. A.1.

A natural measure of the distance from x to the mean μ is the distance $|x-\mu|$ measured in units of standard deviations:

$$r = \frac{|x-\mu|}{\sigma}, \tag{93}$$

MAHALANOBIS DISTANCE

the *Mahalanobis distance* from x to μ. (In the one-dimensional case, this is sometimes called the z-score.) Thus for instance the probability is 0.95 that the Mahalanobis distance from x to μ will be less than 2. If a random variable x is modified by (a) subtracting its mean and (b) dividing by its standard deviation, it is said to be *standardized*. Clearly, a standardized normal random variable $u = (x-\mu)/\sigma$ has zero mean and unit standard deviation—that is,

STANDARDIZED

$$p(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \tag{94}$$

which can be written as $p(u) \sim N(0, 1)$. Table A.1 shows the probability that a value, chosen at random according to $p(u) \sim N(0, 1)$, differs from the mean value by less than a criterion z.

Table A.1. The Probability a Sample Drawn from a Standardized Gaussian has Absolute Value Less Than a Criterion (i.e., $\Pr[|u| \leq z]$)

z	$\Pr[\lvert u\rvert \leq z]$	z	$\Pr[\lvert u\rvert \leq z]$	z	$\Pr[\lvert u\rvert \leq z]$
0.0	0.0	1.0	0.683	2.0	0.954
0.1	0.080	1.1	0.729	2.1	0.964
0.2	0.158	1.2	0.770	2.326	0.980
0.3	0.236	1.3	0.806	2.5	0.989
0.4	0.311	1.4	0.838	2.576	0.990
0.5	0.383	1.5	0.866	3.0	0.9974
0.6	0.452	1.6	0.890	3.090	0.9980
0.7	0.516	1.7	0.911	3.291	0.999
0.8	0.576	1.8	0.928	3.5	0.9995
0.9	0.632	1.9	0.943	4.0	0.99994

A.5 GAUSSIAN DERIVATIVES AND INTEGRALS

Because of the prevalence of Gaussian functions throughout statistical pattern recognition, we often have occasion to integrate and differentiate them. The first three derivatives of a one-dimensional (standardized) Gaussian are

$$\begin{aligned} \frac{\partial}{\partial x}\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}\right] &= \frac{-x}{\sqrt{2\pi}\sigma^3} e^{-x^2/(2\sigma^2)} &&= \frac{-x}{\sigma^2} p(x) \\ \frac{\partial^2}{\partial x^2}\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}\right] &= \frac{1}{\sqrt{2\pi}\sigma^5}\left(-\sigma^2 + x^2\right) e^{-x^2/(2\sigma^2)} &&= \frac{-\sigma^2 + x^2}{\sigma^4} p(x) \\ \frac{\partial^3}{\partial x^3}\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}\right] &= \frac{1}{\sqrt{2\pi}\sigma^7}\left(3x\sigma^2 - x^3\right) e^{-x^2/(2\sigma^2)} &&= \frac{-3x\sigma^2 - x^3}{\sigma^6} p(x), \end{aligned} \tag{95}$$

and are shown in Fig. A.2.

ERROR FUNCTION

An important finite integral of the Gaussian is the so-called *error function*, defined as

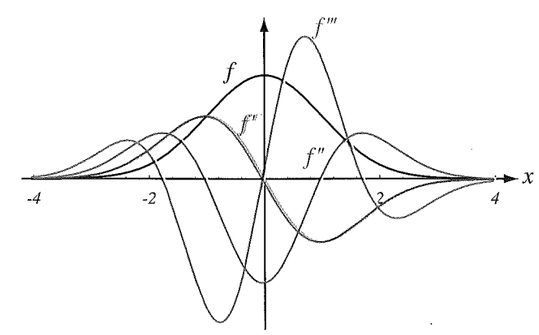

FIGURE A.2. A one-dimensional Gaussian distribution and its first three derivatives, shown for $f(x) \sim N(0, 1)$.

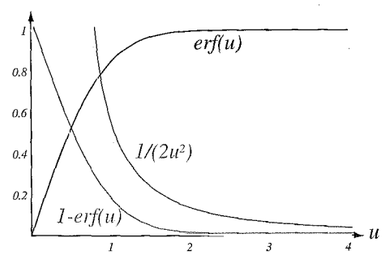

FIGURE A.3. The error function erf(u) corresponds to the area under a standardized Gaussian between $-\sqrt{2}u$ and $\sqrt{2}u$, that is, if x is a standardized Gaussian random variable, $\Pr[|x| \leq \sqrt{2}u] = \text{erf}(u)$. Thus, the complementary probability, $1 - \text{erf}(u)$, is the probability that a sample is chosen with $|x| > \sqrt{2}u$. Chebyshev's inequality states that for an arbitrary distribution having zero mean and unit standard deviation, $\Pr[|x| > \epsilon]$ is less than $1/\epsilon^2$, so that the lower curve is bounded by $1/(2u^2)$. As shown, this bound is quite loose for a Gaussian.

$$\text{erf}(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2} dx. \tag{96}$$

As can be seen from Fig. A.1, erf(0) = 0, and erf(1) = 0.84. There is no closed analytic form for the error function, and thus we typically use tables, approximations, or numerical integration for its evaluation (Fig. A.3).

GAMMA FUNCTION

In calculating moments of Gaussians, we need the general integral of powers of x weighted by a Gaussian. Recall first the definition of a *gamma function*

$$\Gamma(n+1) = \int_0^\infty x^n e^{-x} dx, \tag{97}$$

where the gamma function obeys

$$\Gamma(n) = n\Gamma(n-1) \tag{98}$$

FACTORIAL

and $\Gamma(1/2) = \sqrt{\pi}$. For n an integer we have $\Gamma(n+1) = n \times (n-1) \times (n-2) \cdots \times 1 = n!$, read "n factorial."

Changing variables in Eq. 97, we find the moments of a (normalized) Gaussian distribution as

$$2\int_0^\infty x^n \frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma} dx = \frac{2^{n/2}\sigma^n}{\sqrt{\pi}} \Gamma\left(\frac{n+1}{2}\right), \tag{99}$$

where again we have used a prefactor of 2 and lower integration limit of 0 in order give nontrivial (i.e., nonvanishing) results for odd n.

A.5.1 Multivariate Normal Densities

Normal random variables have many desirable theoretical properties. For example, it turns out that the convolution of two Gaussian functions is again a Gaussian function,

and thus the distribution for the sum of two independent normal random variables is again normal. In fact, sums of dependent normal random variables also have normal distributions. Suppose that each of the d random variables x_i is normally distributed, each with its own mean and variance: $p_{x_i}(x_i) \sim N(\mu_i, \sigma_i^2)$. If these variables are independent, their joint density has the form

$$p(\mathbf{x}) = \prod_{i=1}^d p(x_i) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i} e^{-1/2((x_i-\mu_i)/\sigma_i)^2}$$
$$= \frac{1}{(2\pi)^{d/2}\prod_{i=1}^d \sigma_i} \exp\left[-\frac{1}{2}\sum_{i=1}^d \left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]. \tag{100}$$

This can be written in a compact matrix form if we observe that for this case the covariance matrix is diagonal, that is,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{bmatrix}, \tag{101}$$

and hence the inverse of the covariance matrix is easily written as

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_d^2 \end{bmatrix}. \tag{102}$$

Thus, the exponent in Eq. 100 can be rewritten using

$$\sum_{i=1}^d \left(\frac{x_i-\mu_i}{\sigma_i}\right)^2 = (\mathbf{x}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}). \tag{103}$$

Finally, by noting that the determinant of $\mathbf{\Sigma}$ is just the product of the variances, we can write the joint density compactly in terms of the quadratic form

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]. \tag{104}$$

MULTIVARIATE NORMAL DENSITY

This is the general form of a *multivariate normal density function*, where the covariance matrix $\mathbf{\Sigma}$ is no longer required to be diagonal. With a little linear algebra, it can be shown that if $\mathbf{x}$ obeys this probability law, then

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x}$$

$$\mathbf{\Sigma} = \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int_{-\infty}^{\infty} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})\, d\mathbf{x}, \tag{105}$$

just as one would expect. Multivariate normal data tend to cluster about the mean vector, $\boldsymbol{\mu}$, falling in an ellipsoidally shaped cloud whose principal axes are the eigenvectors of the covariance matrix. The natural measure of the distance from $\mathbf{x}$ to the mean $\boldsymbol{\mu}$ is provided by the quantity

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{106}$$

MAHALANOBIS DISTANCE

which is the square of the *Mahalanobis distance* from $\mathbf{x}$ to $\boldsymbol{\mu}$. It is not as easy to standardize a vector random variable (reduce it to zero mean and unit covariance matrix) as it is in the univariate case. The expression analogous to $u = (x - \mu)/\sigma$ is $\mathbf{u} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, which involves the "square root" of the inverse of the covariance matrix. The process of obtaining $\boldsymbol{\Sigma}^{-1/2}$ requires finding the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$, and it is just a bit beyond the scope of this Appendix.

A.5.2 Bivariate Normal Densities

It is illuminating to look at the bivariate normal density—that is, the case of two normally distributed random variables x_1 and x_2. It is convenient to define $\sigma_1^2 = \sigma_{11}$, $\sigma_2^2 = \sigma_{22}$ and to introduce the correlation coefficient ρ defined by

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}. \tag{107}$$

With this notation, the covariance matrix becomes

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{108}$$

and its determinant simplifies to

$$|\boldsymbol{\Sigma}| = \sigma_1^2 \sigma_2^2 (1 - \rho^2). \tag{109}$$

Thus, the inverse covariance matrix is given by

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} &= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} \\ &= \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1\sigma_2) \\ -\rho/(\sigma_1\sigma_2) & 1/\sigma_2^2 \end{bmatrix}. \end{aligned} \tag{110}$$

Next we explicitly expand the quadratic form in the normal density:

$$\begin{aligned} &(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ &= [(x_1 - \mu_1)\ (x_2 - \mu_2)] \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1\sigma_2) \\ -\rho/(\sigma_1\sigma_2) & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\ &= \frac{1}{1-\rho^2} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right]. \end{aligned} \tag{111}$$

Thus, the general bivariate normal density has the form

$$\begin{aligned} p_{x_1x_2}(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\times \exp\left[-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right] \right]. \end{aligned} \tag{112}$$

As we can see from Fig. A.4, $p(x_1, x_2)$ is a hill-shaped surface over the x_1x_2 plane. The peak of the hill occurs at the point $(x_1, x_2) = (\mu_1, \mu_2)$—that is, at the mean vector $\boldsymbol{\mu}$. The shape of the hump depends on the two variances σ_1^2 and σ_2^2, and the correlation coefficient ρ. If we slice the surface with horizontal planes parallel to the x_1x_2 plane, we obtain the so-called *level curves*, defined by the locus of points where the quadratic form

$$\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \tag{113}$$

is constant. It is not hard to show that $|\rho| \leq 1$ and that this implies that the level curves are ellipses. The x and y extent of these ellipses are determined by the variances σ_1^2 and σ_2^2, and their eccentricity is determined by ρ. More specifically, the *principal axes* of the ellipse are in the direction of the eigenvectors $\mathbf{e}_i$ of $\boldsymbol{\Sigma}$, and the different widths in these directions are $\sqrt{\lambda_i}$. For instance, if $\rho = 0$, the principal axes of the ellipses are parallel to the coordinate axes, and the variables are statistically independent. In the special cases where $\rho = 1$ or $\rho = -1$, the ellipses collapse to straight lines. Indeed, the joint density becomes singular in this situation, because there is really only one independent variable. We shall avoid this degeneracy by assuming that $|\rho| < 1$.

PRINCIPAL AXES

One of the important properties of the multivariate normal density is that all conditional and marginal probabilities are also normal. To find such a density explicitly, which we denote $p_{x_2|x_1}(x_2|x_1)$, we substitute our formulas for $p_{x_1x_2}(x_1, x_2)$ and

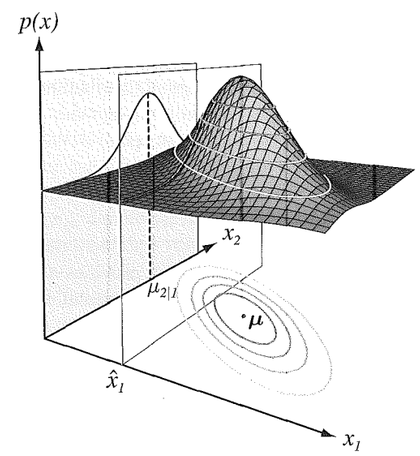

FIGURE A.4. A two-dimensional Gaussian having mean $\boldsymbol{\mu}$ and nondiagonal covariance $\boldsymbol{\Sigma}$. If the value on one variable is known, for instance $x_1 = \hat{x}_1$, the distribution over the other variable is Gaussian with mean $\mu_{2|1}$.

$p_{x_1}(x_1)$ in the defining equation

$$
\begin{aligned}
p_{x_2|x_1}(x_2|x_1) &= \frac{p_{x_1x_2}(x_1, x_2)}{p_{x_1}(x_1)} \\
&= \left[\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}\right] \\
&\quad \times \left[\sqrt{2\pi}\sigma_1 e^{\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2}\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left[\frac{x_2-\mu_2}{\sigma_2} - \rho\frac{x_1-\mu_1}{\sigma_1}\right]^2\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}\left(\frac{x_2 - [\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)]}{\sigma_2\sqrt{1-\rho^2}}\right)^2\right]. \qquad (114)
\end{aligned}
$$

CONDITIONAL MEAN

Thus, we have verified that the conditional density $p_{x_1|x_2}(x_1|x_2)$ is a normal distribution. Moreover, we have explicit formulas for the *conditional mean* $\mu_{2|1}$ and the conditional variance $\sigma^2_{2|1}$:

$$
\mu_{2|1} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1) \qquad \text{and} \qquad \sigma^2_{2|1} = \sigma_2^2(1-\rho^2), \qquad (115)
$$

as illustrated in Fig. A.4.

These formulas provide some insight into the question of how knowledge of the value of x_1 helps us to estimate x_2. Suppose that we know the value of x_1. Then a natural estimate for x_2 is the conditional mean, $\mu_{2|1}$. In general, $\mu_{2|1}$ is a linear function of x_1; if the correlation coefficient ρ is positive, the larger the value of x_1, the larger the value of $\mu_{2|1}$. If it happens that x_1 is the mean value μ_1, then the best we can do is to guess that x_2 is equal to μ_2. Also, if there is no correlation between x_1 and x_2, we ignore the value of x_1, whatever it is, and we always estimate x_2 by μ_2. Note that in that case the variance of x_2, given that we know x_1, is the same as the variance for the marginal distribution, that is, $\sigma^2_{2|1} = \sigma_2^2$. If there is correlation, knowledge of the value of x_1, whatever the value is, reduces the variance. Indeed, with 100% correlation there is no variance left in x_2 when the value of x_1 is known.

A.6 HYPOTHESIS TESTING

Statistical hypothesis testing provides a formal way to decide if the results of an experiment are significant or accidental. It is standard statistical terminology to call a set of n measurements $\mathcal{X}_n = \{x_1, x_2, \ldots, x_n\}$ a *sample of size n*. However, in keeping with the terminology that is universally used in pattern recognition, we shall call each individual measurement a sample. Suppose that we have a set of samples that are drawn either from a known distribution D_0 or from some other distribution. In pattern classification, we seek to determine which distribution was the source of any sample; and if it is indeed D_0, we would classify the point accordingly. Hypothesis testing addresses a somewhat different but related problem. We assume initially that distribution D_0 is the source of the patterns; this is called the *null hypothesis* and is often denoted H_0. Based on the value of any observed sample, we ask whether we can

reject the null hypothesis—that is, state with some degree of confidence (expressed as a probability) that the sample did *not* come from D_0.

For instance, D_0 might be a standardized Gaussian, $p(x) \sim N(0, 1)$, and hence our null hypothesis is that a sample comes from a Gaussian with mean $\mu = 0$. If the value of a particular sample is small (e.g., $x = 0.3$), it is likely that it came from the D_0; after all, 68% of the samples drawn from that distribution have absolute value less than $x = 1.0$ (cf. Fig. A.1). If a sample's value is large (e.g., $x = 5$), then we would be more confident that it did *not* come from D_0. At such a situation we merely conclude that (with some probability) the sample was drawn from a distribution with $\mu \neq 0$.

STATISTICAL SIGNIFICANCE

Viewed another way, for any confidence—expressed as a probability—there exists a criterion value such that if the sampled value differs from $\mu = 0$ by more than that criterion, we reject the null hypothesis. (It is traditional to use confidences of .01 or .05.) We then say that the difference of the sample from 0 is *statistically significant*. For instance, if our null hypothesis is a standardized Gaussian, then if our sample differs from the value $x = 0$ by more than 2.576, we could reject the null hypothesis "at the .01 confidence level," as can be deduced from Table A.1. A more sophisticated analysis could be applied if *several* samples are all drawn from D_0 or if the null hypothesis involved a distribution other than a Gaussian. Of course, this usage of "significance" applies only to the statistical properties of the problem—it implies nothing about whether the results are "important." Hypothesis testing is of great generality, and it is useful when we seek to know whether something other than the assumed case (the null hypothesis) is likely to be the case.

A.6.1 Chi-Squared Test

Hypothesis testing can be applied to discrete problems too. Suppose we have n patterns—n_1 of which are known to be in ω_1, and n_2 in ω_2—and we are interested in determining whether a particular decision rule is useful or informative. In this case, the null hypothesis is that a random decision rule is present—one that selects a pattern and with some probability P places it in a category which we will call the "left" category, and otherwise in the "right" category. We say that a candidate rule is informative if it differs signficantly from such a random decision.

What we need is a clear mathematical definition of statistical significance under these conditions. The random rule (the null hypothesis) would place Pn_1 patterns from ω_1 and Pn_2 from ω_2 independently in the left category and the remainder in the right category. Our candidate decision rule would differ significantly from the random rule if the proportions differed significantly from those given by the random rule. Formally, we let n_{iL} denote the number of patterns from category ω_i placed in the left category by our candidate rule. The so-called *chi-squared* statistic for this case is

$$
\chi^2 = \sum_{i=1}^{2} \frac{(n_{iL} - n_{ie})^2}{n_{ie}}, \qquad (116)
$$

where, according to the null hypothesis, the number of patterns in category ω_i that we expect to be placed in the left category is $n_{ie} = Pn_i$. Clearly χ^2 is nonnegative, and it is zero if and only if all the observed numbers n_{iL} match the expected numbers n_{ie}. The higher the value of χ^2, the less likely it is that the null hypothesis is true. Thus, for a sufficiently high χ^2, the difference between the expected and observed distributions is statistically significant, we can reject the null hypothesis, and we