8

Stochastic Descriptions of Objects and Images

There are many random, unpredictable physical effects that influence the structure of images. The inherent randomness that occurs in photoelectric detection and the noise limits imposed by basic thermodynamics inevitably make images *noisy* or *stochastic* (Greek *stochos*, aim, guess, chance). Additional randomness can arise from a variety of mechanisms in real image detectors. A full description of imaging systems requires analysis of all of these processes. Moreover, any imaging system will be used for a variety of objects, and the randomness of the objects themselves must be taken into account for many purposes.

The natural stochastic description for a digital image is as a finite-dimensional random vector, where each component corresponds to the gray value of a single pixel or to an individual measurement. Objects, on the other hand, are more accurately described as functions of continuous spatial or temporal variables (hence as vectors in an infinite-dimensional vector space); when these functions are stochastic in nature, they are called random processes. In either case, a *stochastic model* is at least a partial description of the statistics of the random vector or process.

Stochastic models have many uses in image science. They are needed for computing simple statistical descriptors such as moments and autocorrelation functions; they allow realistic computer simulation of typical images, and they provide the framework for pattern recognition, image analysis and data compression. In image reconstruction, it is useful to incorporate prior information about the object, and this information is often statistical in nature. Furthermore, as we shall see in detail in Chap. 14, objective assessment of image quality necessarily requires knowledge of the statistical properties of images, and these in turn are sensitive to the statistical properties of the objects being imaged.

It is our objective in this chapter to lay the groundwork for discussing all of these manifestations of randomness. As a starting point, we assume that the reader has a good grasp of the basic concepts of probability and random variables as surveyed in App. C. In Sec. 8.1 we discuss multivariate probability and vector random variables in general terms, though without reference to specific probability laws. Random processes are treated in similar generality in Sec. 8.2. In Sec. 8.3 we discuss an important class of specific probability laws, the Gaussian or normal distributions, as applied to random vectors and random processes. In Sec. 8.4, we introduce a few of the many stochastic models that have been used for random objects, and in Sec. 8.5 we extend the discussion to images (as opposed to objects).

A notable omission in this chapter is any discussion of the Poisson distribution, which plays a crucial role in stochastic modeling of many imaging systems; that omission will be remedied in Chap. 11.

The assistance of Robert F. Wagner in formulating and writing this chapter is gratefully acknowledged.

8.1 RANDOM VECTORS

In Sec. C.2.1 of App. C, a random variable was defined as a function that maps the sample space S of some experiment onto the set of real numbers. That is, each experimental outcome ζ in S is associated with a real scalar $g(\zeta)$. To generalize this idea to a random vector, we need only consider a vector-valued function $\mathbf{g}(\zeta)$.

For example, suppose we want to measure the irradiance of a light beam at some location. We can insert an appropriate photodetector at that location, and the detector output is a scalar random variable. If the beam consists of white light, however, we might want to know the irradiance in each of three color bands. In that case we can use three photodetectors and an arrangement of beamsplitters and filters so that each measurement yields three scalars, which we can regard as components of a three-dimensional (3D) random vector.

Repeated scalar measurements can also be arranged as a vector. If we measure the irradiance at some location K times with a single photodetector, it is often useful to think of the result as a K-dimensional (KD) random vector. Alternatively, we might be interested in the spatial distribution of light in some image plane. We can use an array of M photodetectors and measure the irradiance at M different locations simultaneously, regarding the result as an MD random vector.

Finally, complex scalar random variables can be regarded as 2D vectors. If we measure the amplitude A and phase ϕ of an electromagnetic wave received on an antenna, these quantities can be regarded as two components of a vector. We can also use A and ϕ to compute the real and imaginary parts of a complex number $g = g' + ig''$, and the components g' and g'' are naturally depicted as Cartesian coordinates of a random vector in the complex plane. Equivalently, we can think of g as a complex random scalar if that is convenient. If we measure M complex numbers, we can display the results as either a vector with M complex components or one with $2M$ real components.

It is the goal of this section to establish notation and procedures for dealing with all of these manifestations of random vectors.

8.1.1 Basic concepts

A real MD random vector, denoted $\mathbf{g}$, can be formed from any collection of M real scalar random variables $\{g_m, m = 1, ..., M\}$. For definiteness, the elements will be arranged as a column, so $\mathbf{g}$ is a column vector or $M \times 1$ matrix.

An MD complex random vector $\mathbf{g}$ has components $g_m = g'_m + ig''_m$. It can be represented by the $M \times 1$ column vector of complex random values $(g_1, g_2, ..., g_M)^T$, or as the $2M \times 1$ column vector $(g'_1, g'_2, ..., g'_M, g''_1, g''_2, ..., g''_M)^T$. Hence it is equivalent to think of an MD vector of complex numbers as residing in either $\mathbb{C}^M$ or $\mathbb{R}^{2M}$. Therefore, the treatment in this chapter is often given in terms of real random vectors, with the understanding that the complex case can be obtained by doubling the number of components in the random vector to include both real and imaginary parts as separate elements.

The probability law for a random vector is nothing more than the multivariate probability law for all of its components. Like any other random variable, each component of a random vector is either discrete-valued or continuous-valued. If each component can take on only a finite set of values, or at most a countably infinite set, then we refer to the random vector as discrete-valued. The probability law of a discrete-valued random variable specifies the probability associated with all possible combinations of values for all components. If all of the components of an MD random vector $\mathbf{g}$ are continuous-valued random variables, the full probability law is a multivariate probability density function (PDF) $\mathrm{pr}(g_1, g_2, ..., g_M)$.

The cumulative distribution function for a random vector is defined analogously to that of a scalar random variable [*cf.* (C.26)]:

$$\mathrm{F}(\mathbf{c}) \equiv \Pr(g_1 \leq c_1, g_2 \leq c_2, ..., g_M \leq c_M)\,, \tag{8.1}$$

where $\mathbf{c}$ is a vector with components $\{c_i\}$.

If $\mathbf{g}$ is a continuous-valued random vector, $\mathrm{F}(\mathbf{c})$ is a continuous function of each c_i. Then, in a generalization of (C.29), the PDF on $\mathbf{g}$ can be defined in terms of partial derivatives of $\mathrm{F}(\mathbf{g})$:

$$\mathrm{pr}(\mathbf{g}) \equiv \frac{\partial^M \mathrm{F}(\mathbf{g})}{\partial g_1 \partial g_2 \cdots \partial g_M}\,. \tag{8.2}$$

If we integrate (8.2) we retrieve the cumulative distribution function:

$$\mathrm{F}(\mathbf{g}) = \int_{-\infty}^{g_1} dg'_1 \int_{-\infty}^{g_2} dg'_2 \cdots \int_{-\infty}^{g_M} dg'_M \;\mathrm{pr}(\mathbf{g}')\,. \tag{8.3}$$

A more compact vector notation for (8.3) is

$$\mathrm{F}(\mathbf{g}) = \int_{-\infty}^{\mathbf{g}} d^M g' \;\mathrm{pr}(\mathbf{g}')\,. \tag{8.4}$$

The corresponding expression for a discrete-valued random vector would involve multiple sums in place of the continuous integrals in (8.4), one for each of the components of $\mathbf{g}$.

Marginal probability densities We are often interested in the statistical behavior of a subset of the components of a random vector regardless of the behavior of the others. The statistical description of a single component g_m of the random vector $\mathbf{g}$ is called the marginal probability density function on g_m. To determine the marginal probability density of g_m, we integrate the joint density of $\mathbf{g}$ over all other components:

$$\mathrm{pr}(g_m) = \int_{-\infty}^{\infty} dg_1 \cdots \int_{-\infty}^{\infty} dg_{m-1} \int_{-\infty}^{\infty} dg_{m+1} \cdots \int_{-\infty}^{\infty} dg_M \;\mathrm{pr}(\mathbf{g})\,. \tag{8.5}$$

We can also determine the marginal density of the $(M-1)$-dimensional subvector $\mathbf{g}' = (g_1, g_2, ..., g_{M-1})^t$, which is given by

$$\text{pr}(\mathbf{g}') = \int_{-\infty}^{\infty} dg_M \ \text{pr}(\mathbf{g}) \,. \tag{8.6}$$

Equation (8.5) gives the marginal density of one component of the random vector $\mathbf{g}$; (8.6) gives the marginal density of the random vector $\mathbf{g}'$ formed from all but one component of the random vector $\mathbf{g}$. The marginal density of any other subset of the components of $\mathbf{g}$ is similarly obtained by integrating over all variables not included in the subset.

A simple geometric construction can be used to visualize computation of a marginal. If we compute $\text{pr}(x_0)$ by integrating $\text{pr}(x_0, y)$ over y, we can write that integral as

$$\text{pr}(x_0) = \int_{\infty} dx \int_{\infty} dy \ \text{pr}(x, y)\, \delta(x - x_0) \,. \tag{8.7}$$

The delta function is nonzero on a line parallel to the y-axis, and only values of $\text{pr}(x, y)$ along that line contribute to the integral. The PDF of x_0 is essentially a 1D projection of the 2D joint PDF on (x, y).

Conditional probability densities All of the relations given in Sec. C.4 for joint and conditional probabilities and densities hold for random vectors with minor notational changes. For example, given two random vectors $\mathbf{f}$ and $\mathbf{g}$, Bayes' rule [*cf.* (C.17)] becomes

$$\text{pr}(\mathbf{g}|\mathbf{f}) = \frac{\text{pr}(\mathbf{f}|\mathbf{g})\,\text{pr}(\mathbf{g})}{\text{pr}(\mathbf{f})} \,. \tag{8.8}$$

Two random vectors $\mathbf{f}$ and $\mathbf{g}$ are *statistically independent* if the value of one of them has no influence on the other, that is, $\text{pr}(\mathbf{f}|\mathbf{g}) = \text{pr}(\mathbf{f})$. When two random vectors are independent, their joint PDF factors:

$$\text{pr}(\mathbf{f}, \mathbf{g}) = \text{pr}(\mathbf{g})\,\text{pr}(\mathbf{f}) \,. \tag{8.9}$$

It can be shown that the cumulative distribution function of two independent random vectors also factors.

8.1.2 Expectations

Discrete-valued random vectors Expectation values of discrete-valued random vectors are defined by summing over the possible combinations of the components weighted by the corresponding probabilities. Consider, for example, the MD vector $\mathbf{g}$, where each component g_m can take on any of J values x_j, $j = 1, ..., J$. By extension of the discussion in Sec. C.4, the expectation of an arbitrary function of the components is given by

$$\begin{aligned} &\langle h(g_1, g_2, ..., g_M) \rangle \\ &= \sum_{j_1=1}^{J} \sum_{j_2=1}^{J} \cdots \sum_{j_M=1}^{J} h(x_{j_1}, x_{j_2}, ..., x_{j_M})\, \Pr(g_1 = x_{j_1}, g_2 = x_{j_2}, ..., g_M = x_{j_M}) \,. \end{aligned} \tag{8.10}$$

This notation is cumbersome, but we can shorten it to

$$\langle h(\mathbf{g})\rangle = \sum_{g_1}\sum_{g_2}\cdots\sum_{g_M} h(\mathbf{g})\,\Pr(\mathbf{g})\,, \tag{8.11}$$

where it is understood that each sum runs over the possible values of the component. An even more compact notation with the same meaning is

$$\langle h(\mathbf{g})\rangle = \sum_{\mathbf{g}} h(\mathbf{g})\,\Pr(\mathbf{g})\,, \tag{8.12}$$

where the sum over a vector index signifies a multiple sum over all components running over all possible values.

As in App. C, we shall use the notations $\langle h(\mathbf{g})\rangle$ and $\mathrm{E}\{h(\mathbf{g})\}$ interchangeably, and we shall also use an overbar to denote expectation. Thus, $\overline{\mathbf{g}} = \langle \mathbf{g}\rangle = \mathrm{E}\{\mathbf{g}\}$.

Continuous-valued random vectors Given a continuous-valued random vector $\mathbf{g}$, the expectation of an arbitrary function $h(\mathbf{g})$ is written as

$$\langle h(g_1, g_2, ..., g_M)\rangle = \int_{-\infty}^{\infty} dg_1 \int_{-\infty}^{\infty} dg_2 \cdots \int_{-\infty}^{\infty} dg_M \; h(g_1, g_2, ..., g_M)\,\mathrm{pr}(g_1, g_2, ..., g_M)\,. \tag{8.13}$$

There is no loss of generality in the infinite limits since the density might be zero except on a finite support. In more compact notation, (8.13) becomes

$$\langle h(\mathbf{g})\rangle = \int_{\infty} d^M g \; h(\mathbf{g})\,\mathrm{pr}(\mathbf{g})\,, \tag{8.14}$$

where the subscript ∞ on the integral sign indicates that it runs over an infinite range for each of the M variables of integration.

We have not specified the nature of the function $h(\mathbf{g})$. It could be a scalar-valued or a vector-valued function of the random vector $\mathbf{g}$. It could even be $\mathbf{g}$ itself, in which case $\langle \mathbf{g}\rangle$ is the *mean vector* $\overline{\mathbf{g}}$. The components of this vector are given by

$$\overline{g}_m = \langle g_m\rangle\ . \tag{8.15}$$

For complex vectors, the mean is defined separately for real and imaginary parts. Thus $\mathbf{g} = \mathbf{g}' + i\mathbf{g}''$ implies $g_m = g'_m + i g''_m$ and $\overline{\mathbf{g}} = \overline{\mathbf{g}}' + i\overline{\mathbf{g}}''$, which means that $\overline{g}_m = \overline{g}'_m + i\overline{g}''_m$ for all m.

8.1.3 Covariance and correlation matrices

It is often of interest to know whether two different components of a random vector *covary*, that is, whether fluctuations in one are statistically related to fluctuations in the other. To quantify this concept, we define the *covariance matrix* $\mathbf{K}$. For an MD random vector $\mathbf{g}$, $\mathbf{K}$ is an $M \times M$ matrix with elements given by

$$K_{ij} = \left\langle (g_i - \overline{g}_i)(g_j - \overline{g}_j)^* \right\rangle\ , \tag{8.16}$$

where the asterisk indicates complex conjugate, allowing for the possibility that components of $\mathbf{g}$ might be complex. It is easy to see from this definition that $\mathbf{K}$ is Hermitian, *i.e.*, $K_{ij} = K_{ji}^*$.

For the special case where g_i and g_j are statistically independent, we can write

$$K_{ij} = \langle (g_i - \overline{g}_i) \rangle \left\langle (g_j - \overline{g}_j)^* \right\rangle = 0\,, \qquad i \neq j\,. \tag{8.17}$$

Any random variable covaries with itself. The diagonal elements of the covariance matrix are the variances of the components:

$$K_{jj} = \mathrm{Var}\{g_j\}\,. \tag{8.18}$$

Another way of expressing the covariance matrix is as an *outer product*, as discussed in Sec. 1.3.7. With the notation of (1.53), (8.16) is equivalent to

$$\mathbf{K} = \left\langle (\mathbf{g} - \overline{\mathbf{g}})(\mathbf{g} - \overline{\mathbf{g}})^\dagger \right\rangle = \left\langle \Delta\mathbf{g}\Delta\mathbf{g}^\dagger \right\rangle\,, \tag{8.19}$$

where $\Delta\mathbf{g} \equiv \mathbf{g} - \overline{\mathbf{g}}$.

A related matrix is the *correlation matrix* $\mathbf{R}$, defined as

$$\mathbf{R} = \left\langle \mathbf{g}\,\mathbf{g}^\dagger \right\rangle\,. \tag{8.20}$$

By unscrambling the outer-product notation, we see that $R_{ij} = \left\langle g_i g_j^* \right\rangle$, so $\mathbf{R}$ is the matrix organization of the second moments of the random vector. As a generalization of a well-known relation for two random variables, (C.85), we have

$$\mathbf{K} = \mathbf{R} - \overline{\mathbf{g}}\,\overline{\mathbf{g}}^\dagger\,. \tag{8.21}$$

For zero-mean random vectors, therefore, $\mathbf{R}$ and $\mathbf{K}$ are identical.

When two or more random vectors are involved in the same problem, we shall add appropriate subscripts to $\mathbf{K}$ and $\mathbf{R}$. Thus $\mathbf{R}_\mathbf{g} = \left\langle \mathbf{g}\,\mathbf{g}^\dagger \right\rangle$ and $\mathbf{R}_\mathbf{f} = \left\langle \mathbf{f}\,\mathbf{f}^\dagger \right\rangle$.

Positive-definiteness Every covariance matrix $\mathbf{K}$ is positive-semidefinite, as defined in Sec. A.8. To demonstrate this point, consider an arbitrary quadratic form as in (A.115):

$$\begin{aligned} Q_\mathbf{K}(\mathbf{x}) = \mathbf{x}^\dagger\mathbf{K}\mathbf{x} = \mathbf{x}^\dagger \left\langle \Delta\mathbf{g}\Delta\mathbf{g}^\dagger \right\rangle \mathbf{x} &= \left\langle \mathbf{x}^\dagger \Delta\mathbf{g}\Delta\mathbf{g}^\dagger \mathbf{x} \right\rangle \\ &= \left\langle \left|\mathbf{x}^\dagger \Delta\mathbf{g}\right|^2 \right\rangle\,, \end{aligned} \tag{8.22}$$

where $\mathbf{x}$ is a nonrandom vector and we have used elementary properties of scalar products and norms from Chap. 1. Since $|\mathbf{x}^\dagger\Delta\mathbf{g}|^2$ is never negative, its expectation is never negative, so the quadratic form $Q_\mathbf{K}(\mathbf{x})$ is never negative and $\mathbf{K}$ is positive-semidefinite (nonnegative-definite) by definition.

Moreover, it is rare that covariance matrices are not strictly positive-definite. From Sec. A.8 we know that an $M \times M$ positive-semidefinite matrix is positive-definite if its rank R equals its dimension M, and from Sec. A.3 we know that the rank is the number of linearly independent rows or columns. Thus the only way we can have $R < M$ is if at least one of the columns of $\mathbf{K}$ can be written as a linear combination of the other columns. One way in which this can happen is if not all components of $\mathbf{g}$ are measured independently, but instead one component is computed as a weighted sum of the others. Barring such unusual circumstances, however, it is reasonable to assume that $R = M$.

Cross-covariance and cross-correlation The cross-covariance matrix and the cross-correlation matrix for two random vectors $\mathbf{g}$ and $\mathbf{f}$ are defined analogously to (8.19) and (8.20), respectively. They are related by an expression analogous to (8.21):

$$\mathbf{R}_{\mathbf{gf}} = \left\langle \mathbf{g}\,\mathbf{f}^{\dagger} \right\rangle = \mathbf{K}_{\mathbf{gf}} + \overline{\mathbf{g}}\,\overline{\mathbf{f}}^{\dagger} \,. \tag{8.23}$$

The random vectors $\mathbf{g}$ and $\mathbf{f}$ are said to be *uncorrelated* if their cross-correlation matrix factors as

$$\mathbf{R}_{\mathbf{gf}} = \langle \mathbf{g} \rangle \langle \mathbf{f}^{\dagger} \rangle = \overline{\mathbf{g}}\,\overline{\mathbf{f}}^{\dagger} \,, \tag{8.24}$$

or, equivalently, if their cross-covariance matrix is identically zero.

Since the PDF of two independent random vectors separates into the product of their individual PDFs, we have the immediate result that independent random vectors are uncorrelated. No general statement can be made to the converse, but we shall see later in this chapter that uncorrelated normally distributed random vectors are statistically independent.

Two random vectors are said to be *orthogonal* if

$$\mathbf{R}_{\mathbf{gf}} = \left\langle \mathbf{g}\,\mathbf{f}^{\dagger} \right\rangle = 0 \,. \tag{8.25}$$

Note that this stochastic definition involves the outer product whereas the deterministic definition of orthogonality of two vectors involves the inner product. From (8.23) we see that if the mean of either $\mathbf{g}$ or $\mathbf{f}$ is zero, the cross-correlation and the cross-covariance matrices are equal; in that case orthogonal random vectors are also uncorrelated.

8.1.4 Characteristic functions

The characteristic function $\psi_{\mathbf{g}}(\boldsymbol{\xi})$ of a random vector can be defined as the natural generalization of the characteristic function of a scalar random variable (see Sec. C.3.3). For a real $M \times 1$ random vector $\mathbf{g}$ (column vector), the characteristic function is defined as

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \left\langle \exp(-2\pi i \boldsymbol{\xi}^t \mathbf{g}) \right\rangle \,, \tag{8.26}$$

where $\boldsymbol{\xi}^t$ is a real $1 \times M$ vector[1] (row vector) and hence $\boldsymbol{\xi}^t\mathbf{g}$ is the scalar product of $\mathbf{g}$ and $\boldsymbol{\xi}$.

For the case of a continuous-valued random vector, $\psi_{\mathbf{g}}(\boldsymbol{\xi})$ can be written as

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \int_{\infty} d^M g \; \mathrm{pr}(\mathbf{g}) \exp(-2\pi i \boldsymbol{\xi}^t \mathbf{g}) \,. \tag{8.27}$$

This integral is the MD Fourier transform of the PDF, so the properties of Fourier transforms from Chap. 3 can be used in its manipulation. In particular, since any PDF is nonnegative and normalized to unity, it is in $\mathbb{L}_1$; thus $\psi_{\mathbf{g}}(\boldsymbol{\xi})$ is finite for all $\boldsymbol{\xi}$, is continuous everywhere, and vanishes at infinity [see (3.65) and (3.66)]. The PDF on $\mathbf{g}$ is given by the inverse Fourier transform of $\psi_{\mathbf{g}}(\boldsymbol{\xi})$:

$$\mathrm{pr}(\mathbf{g}) = \int_{\infty} d^M \xi \;\, \psi_{\mathbf{g}}(\boldsymbol{\xi}) \exp(2\pi i \boldsymbol{\xi}^t \mathbf{g}) \,. \tag{8.28}$$

---

[1] One should not confuse $\boldsymbol{\xi}$ with the x component of spatial frequency, denoted ξ in other chapters. The vector $\boldsymbol{\xi}$ used here is a frequency in the sense that it is a variable in a Fourier transform, but it is not a spatial frequency.

The characteristic function of a random vector is unique, in that two random vectors have the same characteristic function if and only if they have the same probability distribution. And, as in the univariate case, two random vectors are independent if and only if their joint characteristic function can be written as the product of their marginal characteristic functions.

Moments The characteristic function has great utility not only for deriving PDFs but also as a shortcut to obtaining the moments of a random vector. This property follows from the definition (8.26) by expanding the exponential in a power series before taking the expectation value. This leads to a series of terms involving increasingly higher moments of the random variable $\mathbf{g}$. These moments can be isolated by differentiation of the series and then setting $\boldsymbol{\xi} = \mathbf{0}$, where $\mathbf{0}$ is the vector with all elements equal to zero. Alternatively, one can simply differentiate the characteristic function directly. For example, if we take the gradient we obtain (in the notation of Sec. A.9.2)

$$\frac{\partial \psi_{\mathbf{g}}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \left\langle (-2\pi i \mathbf{g}) \exp(-2\pi i \boldsymbol{\xi}^t \mathbf{g}) \right\rangle . \tag{8.29}$$

On setting $\boldsymbol{\xi} = \mathbf{0}$, this yields

$$\langle \mathbf{g} \rangle = (-2\pi i)^{-1} \left[\frac{\partial \psi_{\mathbf{g}}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \right]_{\boldsymbol{\xi}=\mathbf{0}} . \tag{8.30}$$

Differentiating twice yields the second moment:

$$\mathbf{R} = \left\langle \mathbf{g}\,\mathbf{g}^t \right\rangle = (-2\pi i)^{-2} \left[\frac{\partial^2 \psi_{\mathbf{g}}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^t} \right]_{\boldsymbol{\xi}=\mathbf{0}} . \tag{8.31}$$

Higher-order moments can be determined using the following general expression:

$$\mathrm{E}\{ g_1^{k_1} g_2^{k_2} ... g_M^{k_M} \} = (-2\pi i)^{k_1+k_2+\cdots+k_M} \left[\frac{\partial^{k_1+k_2+\cdots+k_M} \psi_{\mathbf{g}}(\boldsymbol{\xi})}{\partial \xi_1^{k_1} \partial \xi_2^{k_2} \cdots \partial \xi_M^{k_M}} \right]_{\boldsymbol{\xi}=\mathbf{0}} . \tag{8.32}$$

Complex random vectors The characteristic function of a complex random vector $\mathbf{g}$ can be written

$$\begin{aligned} \psi_{\mathbf{g}}(\boldsymbol{\xi}) &= \left\langle \exp[-2\pi i \,\mathrm{Re}(\boldsymbol{\xi}^\dagger \mathbf{g})] \right\rangle = \left\langle \exp[-\pi i (\boldsymbol{\xi}^\dagger \mathbf{g} + \mathbf{g}^\dagger \boldsymbol{\xi})] \right\rangle \\ &= \langle \exp\left[-2\pi i (\xi_1' g_1' + \xi_1'' g_1'' + \cdots + \xi_M' g_M' + \xi_M'' g_M'') \right] \rangle , \end{aligned} \tag{8.33}$$

where now $\boldsymbol{\xi}$ is an MD complex vector $\boldsymbol{\xi} = \boldsymbol{\xi}' + i\boldsymbol{\xi}''$.

We see that the scalar product in the exponent of (8.33) is the sum of $2M$ real terms, rather than the M terms of (8.26). Another avenue for obtaining this expression is to make use of the fact that complex vectors can be considered to lie in either $\mathbb{C}^M$ or $\mathbb{R}^{2M}$. Thus we could have chosen to represent the MD complex random vector $\mathbf{g}$ by the $2M$D vector of real components $(g_1', g_2', ..., g_M', g_1'', g_2'', ..., g_M'')^t$ and similarly represent $\boldsymbol{\xi}$ by the vector of real components $(\xi_1', \xi_2', ..., \xi_M', \xi_1'', \xi_2'', ..., \xi_M'')^t$. The use of (8.26) with these real vectors would give an expression for the characteristic function identical to (8.33).

The moments of $\mathbf{g}$ can be determined by differentiation of $\psi(\mathbf{g})$ if we mind the rules for differentiation with respect to complex vectors given in Sec. A.9.5. The mean of the random vector $\mathbf{g}$ is found by taking the derivative of $\psi_{\mathbf{g}}(\boldsymbol{\xi})$ with respect to the complex vector $\boldsymbol{\xi}$:

$$\nabla\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \left[\frac{\partial}{\partial\boldsymbol{\xi}'} + i\frac{\partial}{\partial\boldsymbol{\xi}''}\right]\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \left\langle(-2\pi i\mathbf{g})\exp[-\pi i(\boldsymbol{\xi}^\dagger\mathbf{g} + \mathbf{g}^\dagger\boldsymbol{\xi})]\right\rangle , \tag{8.34}$$

where we have made use of (A.159) and (A.160). When we set $\boldsymbol{\xi}$ to zero we find

$$\langle\mathbf{g}\rangle = (-2\pi i)^{-1}\left[\nabla\psi_{\mathbf{g}}(\boldsymbol{\xi})\right]_{\boldsymbol{\xi}=\mathbf{0}} . \tag{8.35}$$

The second moment is found from the generalized Hessian (A.165):

$$\mathbf{R} = \langle\mathbf{g}\,\mathbf{g}^\dagger\rangle = (-2\pi i)^{-2}\left[\nabla\nabla^\dagger\psi_{\mathbf{g}}(\boldsymbol{\xi})\right]_{\boldsymbol{\xi}=0} . \tag{8.36}$$

Higher-order moments can be derived using successive differentiation, similar to the case of real $\mathbf{g}$.

8.1.5 Transformations of random vectors

Section C.3.1 gives rules for transforming PDFs of scalar random variables. A bivariate extension of these rules is presented in Sec. C.4.5. In this section we extend these rules further so that they apply to random vectors of general dimension. Our treatment is limited to real vectors; the extension to complex vectors can be done by converting the complex vectors to real vectors with double the dimension as described above.

Suppose the random vector $\mathbf{g}$ is related to the random vector $\mathbf{f}$ through the general nonlinear relationship $\mathbf{g} = \boldsymbol{\mathcal{O}}\mathbf{f}$. The mapping from $\mathbf{f}$ to $\mathbf{g}$ is discrete-to-discrete even though the components of the vectors are continuous valued. If we assume that this mapping is differentiable (with respect to the component values) and also one-to-one and onto, then the inverse mapping $\mathbf{f} = \boldsymbol{\mathcal{O}}^{-1}(\mathbf{g})$ exists. The PDF of $\mathbf{g}$ is then obtained from the known PDF of $\mathbf{f}$ by recognizing the equivalence of the probability spaces used to describe random events in terms of either $\mathbf{f}$ or $\mathbf{g}$:

$$\mathrm{pr}_{\mathbf{g}}(\mathbf{g})\,d^N\mathbf{g} = \mathrm{pr}_{\mathbf{f}}(\mathbf{f})\,d^N\mathbf{f}. \tag{8.37}$$

The random vector $\mathbf{g}$ must have the same dimensionality as $\mathbf{f}$ if the mapping from $\mathbf{f}$ to $\mathbf{g}$ is invertible. From (8.37) we obtain

$$\mathrm{pr}_{\mathbf{g}}(\mathbf{g}) = \mathrm{pr}_{\mathbf{f}}(\boldsymbol{\mathcal{O}}^{-1}\mathbf{g})|\det\mathbf{J}| , \tag{8.38}$$

where $\mathbf{J}$ is the Jacobian matrix of partial derivatives relating the components of $\mathbf{f}$ and $\mathbf{g}$ [*cf.* (C.102)]:

$$J_{ij} = \frac{\partial f_i}{\partial g_j} , \tag{8.39}$$

and $|\det\mathbf{J}|$ is the absolute value of its determinant.

Linear transformations If the random vector $\mathbf{g}$ is generated as the output of a linear filter acting on the random vector $\mathbf{f}$, we can characterize the linear transformation by an $M \times N$ matrix $\mathbf{H}$. Then we can write the $M \times 1$ output vector $\mathbf{g}$ in terms of the $N \times 1$ input vector $\mathbf{f}$ as

$$\mathbf{g} = \mathbf{H}\mathbf{f}. \tag{8.40}$$

If $M = N$ and $\mathbf{H}^{-1}$ exists, the PDF of $\mathbf{g}$ can be written in terms of the PDF of $\mathbf{f}$ as a special case of (8.38):

$$\mathrm{pr}_{\mathbf{g}}(\mathbf{g}) = \mathrm{pr}_{\mathbf{f}}(\mathbf{H}^{-1}\mathbf{g})\,|\det \mathbf{H}^{-1}|\,. \tag{8.41}$$

Characteristic function of the transformed vector If $\mathbf{H}$ is not invertible, we cannot use (8.41) to relate the PDF for $\mathbf{g}$ to the PDF for $\mathbf{f}$, but we can relate the corresponding characteristic functions. With (8.40), (8.26) becomes

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \left\langle \exp(-2\pi i \boldsymbol{\xi}^t \mathbf{H}\mathbf{f}) \right\rangle = \left\langle \exp\left[-2\pi i (\mathbf{H}^t \boldsymbol{\xi})^t \mathbf{f}\right] \right\rangle , \tag{8.42}$$

where the last step has used the definition of the adjoint, (1.39). (Since we are considering real matrices here, adjoint is the same as transpose.) Comparison of the last expectation in (8.42) with (8.26) shows that

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \psi_{\mathbf{f}}(\mathbf{H}^t \boldsymbol{\xi})\,, \tag{8.43}$$

so knowledge of $\psi_{\mathbf{f}}$ and $\mathbf{H}$ immediately gives $\psi_{\mathbf{g}}$. As an exercise, the reader can show that (8.43) and (8.38) are equivalent if $\mathbf{H}^{-1}$ exists.

The PDF on $\mathbf{g}$ can in principle be found by taking an inverse MD Fourier transform of (8.43). Formally, we can write

$$\mathrm{pr}(\mathbf{g}) = \int_\infty d^M\xi \;\psi_{\mathbf{f}}(\mathbf{H}^t \boldsymbol{\xi}) \exp(2\pi i \boldsymbol{\xi}^t \mathbf{g})\,, \tag{8.44}$$

but in practice the integral might not be easy. The problem is that we are integrating a function of an ND vector over an MD space.

Alternative approach Another way to derive an expression for the PDF of $\mathbf{g}$, when $\mathbf{g} = \mathbf{H}\mathbf{f}$, is to use the multivariate counterpart of (C.77) to write

$$\mathrm{pr}(\mathbf{g}) = \int_\infty d^N f \;\mathrm{pr}(\mathbf{g}|\mathbf{f})\,\mathrm{pr}(\mathbf{f})\,. \tag{8.45}$$

Here the notation $\mathrm{pr}(\mathbf{g}|\mathbf{f})$ is a bit tricky: $\mathbf{g}$ is *defined* as $\mathbf{H}\mathbf{f}$ (not $\mathbf{H}\mathbf{f} + \mathbf{n}$ here), so once $\mathbf{f}$ is given, $\mathbf{g}$ is no longer random; it is just $\mathbf{H}\mathbf{f}$. Nevertheless, we can still use (8.45) if we let $\mathrm{pr}(\mathbf{g}|\mathbf{f})$ be the MD delta function, $\delta(\mathbf{g} - \mathbf{H}\mathbf{f})$. Then we have

$$\mathrm{pr}(\mathbf{g}) = \int_\infty d^N f \;\delta(\mathbf{g} - \mathbf{H}\mathbf{f})\,\mathrm{pr}(\mathbf{f})\,. \tag{8.46}$$

This form is, in fact, equivalent to (8.42). If we take the MD Fourier transform of both sides of (8.46), we find

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \mathcal{F}_M\{\mathrm{pr}(\mathbf{g})\} = \int_\infty d^M g \int_\infty d^N f \;\delta(\mathbf{g} - \mathbf{H}\mathbf{f})\,\mathrm{pr}(\mathbf{f}) \exp(-2\pi i \boldsymbol{\xi}^t \mathbf{g})\,. \tag{8.47}$$

The delta function allows us to perform the integral over $\mathbf{g}$, and we obtain

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \int_\infty d^N f \;\mathrm{pr}(\mathbf{f}) \exp(-2\pi i \boldsymbol{\xi}^t \mathbf{H}\mathbf{f}) = \left\langle \exp(-2\pi i \boldsymbol{\xi}^t \mathbf{H}\mathbf{f}) \right\rangle . \tag{8.48}$$

This equation is the same as (8.42), and (8.43) follows as before.

Although (8.43) and (8.46) are equivalent, the latter may be easier to interpret geometrically. Suppose $M < N$. Then the integral is over an ND space but the delta function is nonzero only on an MD hyperplane defined by $\mathbf{g} = \mathbf{Hf}$. Only vectors $\mathbf{f}$ that lie on this hyperplane make any contribution to the integral for a particular $\mathbf{g}$. This is similar to the geometric construction we presented for the computation of a marginal in (8.7).

Transformation of the mean and covariance When $\mathbf{g} = \mathbf{Hf}$, all moments of $\mathbf{g}$ can be derived by differentiating (8.43), but often we are interested in just the mean or covariance matrix. From the linearity of the expectation operator, we have immediately for the mean of $\mathbf{g}$,

$$\overline{\mathbf{g}} = \langle \mathbf{g} \rangle = \langle \mathbf{Hf} \rangle = \mathbf{H} \langle \mathbf{f} \rangle = \mathbf{H}\overline{\mathbf{f}}. \tag{8.49}$$

The covariance matrix of $\mathbf{g}$ is found as

$$\mathbf{K_g} = \left\langle \Delta\mathbf{g}\Delta\mathbf{g}^\dagger \right\rangle = \left\langle (\mathbf{Hf} - \mathbf{H}\overline{\mathbf{f}})(\mathbf{Hf} - \mathbf{H}\overline{\mathbf{f}})^\dagger \right\rangle = \mathbf{H} \left\langle \Delta\mathbf{f}\Delta\mathbf{f}^\dagger \right\rangle \mathbf{H}^\dagger = \mathbf{H}\mathbf{K_f}\mathbf{H}^\dagger \,, \tag{8.50}$$

where $\Delta\mathbf{f} \equiv \mathbf{f} - \overline{\mathbf{f}}$. The same results can, of course, also be found from (8.43).

These rules for transforming means and covariance matrices will recur often in this book.

8.1.6 Eigenanalysis of covariance matrices

A covariance matrix is Hermitian, and we saw in Sec. 1.4.4 that eigenvectors and eigenvalues of Hermitian matrices have many nice properties. The eigenvalues are real, and the eigenvectors can be chosen to form a complete, orthonormal set in the domain of the matrix. Expansion of a random vector in eigenvectors of its covariance matrix is a valuable tool in statistical analysis.

Let $\mathbf{K_g}$ be the $M \times M$ covariance matrix for a random vector $\mathbf{g}$. The eigenvalue equation for this matrix is

$$\mathbf{K_g}\boldsymbol{\phi}_m = \mu_m \boldsymbol{\phi}_m \,, \qquad m = 1, ..., M \,, \tag{8.51}$$

where $\boldsymbol{\phi}_m$ is an $M \times 1$ eigenvector and μ_m is the corresponding eigenvalue. (Note that the subscript on $\boldsymbol{\phi}_m$ denotes a particular eigenvector, not a component.) Since $\mathbf{K_g}$ is Hermitian, μ_m is real even if $\mathbf{K_g}$ is complex.

We showed above that $\mathbf{K_g}$ is at least positive-semidefinite, so $\mu_m \geq 0$ for all m. For convenience we assume that the eigenvalues are labeled by decreasing value:

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_R > 0 \,, \tag{8.52}$$

where R is the rank. We know from Sec. 1.4.3 that the rank is also the number of nonzero eigenvalues, so μ_R is the smallest nonzero eigenvalue. We also argued above that the rank of $\mathbf{K_g}$ is likely to be the dimension M, in which case there are no nonzero eigenvalues. Then $\mathbf{K_g}$ is positive-definite and hence nonsingular (see Sec. 1.4.3).

We know from Sec. 1.4.4 that the eigenvectors of $\mathbf{K_g}$ can always be chosen as a complete, orthonormal set. The orthonormality can be expressed in inner-product notation as

$$\boldsymbol{\phi}_m^\dagger \boldsymbol{\phi}_n = \delta_{mn} \,, \tag{8.53}$$

where $\boldsymbol{\phi}_m^\dagger$ is the row vector obtained by transposing the column vector $\boldsymbol{\phi}_m$ and taking an element-by-element complex conjugate. The completeness of the eigenvectors is expressed by the closure relation,

$$\sum_{m=1}^{M} \boldsymbol{\phi}_m \boldsymbol{\phi}_m^\dagger = \mathbf{I}\,, \tag{8.54}$$

where $\boldsymbol{\phi}_m \boldsymbol{\phi}_m^\dagger$ is an outer product (see Sec. 1.3.7) and $\mathbf{I}$ is the $M \times M$ unit matrix.

From the discussion in Sec. 1.4.5, we know that the eigenvalue problem (8.51) can also be expressed as

$$\mathbf{K_g}\boldsymbol{\Phi} = \boldsymbol{\Phi}\mathbf{M}\,, \tag{8.55}$$

where $\boldsymbol{\Phi}$ is a matrix formed by arraying the column vectors $\boldsymbol{\phi}_m$ side by side and $\mathbf{M}$ is a diagonal matrix with the m^{th} diagonal element equal to μ_m. (Note that $\mathbf{M}$ is capital μ.) From (8.53) and (8.54), it follows that $\boldsymbol{\Phi}$ is a unitary matrix, *i.e.*, $\boldsymbol{\Phi}^{-1} = \boldsymbol{\Phi}^\dagger$. From this property, we immediately find a useful representation of the covariance matrix [*cf.* (1.85)]:

$$\mathbf{K_g} = \boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Phi}^\dagger\,. \tag{8.56}$$

This representation can also be expressed in terms of outer products [*cf.* (1.86)] as

$$\mathbf{K_g} = \sum_{m=1}^{M} \mu_m \boldsymbol{\phi}_m \boldsymbol{\phi}_m^\dagger\,. \tag{8.57}$$

This expression is the *spectral decomposition* of the covariance matrix.

Discrete Karhunen-Loève expansion Since the eigenvectors of a Hermitian operator form a complete, orthonormal set in the relevant space, any $M \times 1$ vector $\mathbf{g}$ can be expressed as

$$\mathbf{g} = \sum_{m=1}^{M} \beta_m \boldsymbol{\phi}_m\,, \tag{8.58}$$

where the coefficients are given by

$$\beta_m = \boldsymbol{\phi}_m^\dagger \mathbf{g}\,. \tag{8.59}$$

We can express these relations in matrix-vector form by defining an $M \times 1$ vector $\boldsymbol{\beta}$ with components $\{\beta_m\}$. Then

$$\mathbf{g} = \boldsymbol{\Phi}\boldsymbol{\beta}\,, \qquad \boldsymbol{\beta} = \boldsymbol{\Phi}^\dagger \mathbf{g}\,. \tag{8.60}$$

These equations are quite general, holding for any $\mathbf{g}$ and any orthonormal basis vectors. If, however, $\mathbf{g}$ is a random vector and the vectors $\{\boldsymbol{\phi}_m\}$ are eigenvectors of its covariance matrix, then the coefficients $\{\beta_m\}$ are uncorrelated random variables. It is easy to demonstrate this point. In component form, the covariance matrix for $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\langle \Delta\beta_n \Delta\beta_m^* \rangle &= \left\langle \left[\boldsymbol{\phi}_n^\dagger \Delta\mathbf{g}\right] \left[\boldsymbol{\phi}_m^\dagger \Delta\mathbf{g}\right]^* \right\rangle = \left\langle \boldsymbol{\phi}_n^\dagger \Delta\mathbf{g}\Delta\mathbf{g}^\dagger \boldsymbol{\phi}_m \right\rangle \\ &= \boldsymbol{\phi}_n^\dagger \left\langle \Delta\mathbf{g}\Delta\mathbf{g}^\dagger \right\rangle \boldsymbol{\phi}_m = \boldsymbol{\phi}_n^\dagger \mathbf{K_g} \boldsymbol{\phi}_m = \mu_m \boldsymbol{\phi}_n^\dagger \boldsymbol{\phi}_m = \mu_m\, \delta_{nm}\,,\end{aligned} \tag{8.61}$$

where $\Delta\beta_m \equiv \beta_m - \langle\beta_m\rangle$ and we have used the eigenvalue equation (8.51) and the orthonormality of the eigenvectors (8.53). In matrix form, (8.61) reads

$$\mathbf{K}_{\boldsymbol{\beta}} = \boldsymbol{\Phi}^\dagger \mathbf{K}_\mathbf{g} \boldsymbol{\Phi} = \boldsymbol{\Phi}^\dagger \boldsymbol{\Phi} \mathbf{M} \boldsymbol{\Phi}^\dagger \boldsymbol{\Phi} = \mathbf{M}\,, \tag{8.62}$$

where we have used (8.56), (8.60) and the unitarity of $\boldsymbol{\Phi}$.

Expansion of a random vector in eigenvectors of its covariance matrix is known as *Karhunen-Loève* or KL expansion. The key feature of a KL expansion is that the coefficients are uncorrelated (since $\mathbf{K}_{\boldsymbol{\beta}}$ is diagonal). A similar expansion for random processes will be presented in Sec. 8.2.7.

The KL expansion enables us immediately to find a useful representation of the inverse of a covariance matrix. Since $\boldsymbol{\Phi}$ is a unitary matrix, *i.e.*, $\boldsymbol{\Phi}^{-1} = \boldsymbol{\Phi}^\dagger$, we can use (8.62) to write the covariance matrix $\mathbf{K}_\mathbf{g}$ as

$$\mathbf{K}_\mathbf{g} = \boldsymbol{\Phi} \mathbf{M} \boldsymbol{\Phi}^\dagger\,. \tag{8.63}$$

The inverse of $\mathbf{K}_\mathbf{g}$ is then given by

$$\mathbf{K}_\mathbf{g}^{-1} = \boldsymbol{\Phi} \mathbf{M}^{-1} \boldsymbol{\Phi}^\dagger\,, \tag{8.64}$$

where $\mathbf{M}^{-1}$ is also diagonal, with the m^{th} diagonal element given by $1/\mu_m$. Thus the same matrix that diagonalizes $\mathbf{K}_\mathbf{g}$ also diagonalizes $\mathbf{K}_\mathbf{g}^{-1}$.

Whitening As we have just seen, the KL expansion results in a vector $\boldsymbol{\beta}$ with uncorrelated components. It is often useful to go further and force the components all to have the same variance. The concept of a square-root matrix, discussed in Sec. A.8.3, provides us with the tool to accomplish this goal.

By analogy to (A.118), we can define the square root of the covariance matrix of $\mathbf{g}$ by

$$\mathbf{K}_\mathbf{g}^{\frac{1}{2}} = \sum_{m=1}^{M} \sqrt{\mu_m}\, \boldsymbol{\phi}_m \boldsymbol{\phi}_m^\dagger\,. \tag{8.65}$$

If $\mathbf{K}_\mathbf{g}$ is nonsingular, as it usually is, we can write the inverse of the square-root matrix as

$$\mathbf{K}_\mathbf{g}^{-\frac{1}{2}} = \sum_{m=1}^{M} \frac{1}{\sqrt{\mu_m}} \boldsymbol{\phi}_m \boldsymbol{\phi}_m^\dagger\,. \tag{8.66}$$

To verify that this is the correct form for the inverse, one can multiply (8.65) by (8.66) and use the orthonormality relation (8.53) to obtain (8.54).

We now define the vector $\mathbf{y}$ by

$$\mathbf{y} = \mathbf{K}_\mathbf{g}^{-\frac{1}{2}} (\mathbf{g} - \overline{\mathbf{g}})\,. \tag{8.67}$$

With this construction $\overline{\mathbf{y}} = \mathbf{0}$, and its covariance matrix is given by

$$\mathbf{K}_\mathbf{y} = \left\langle \mathbf{y}\mathbf{y}^\dagger \right\rangle = \mathbf{K}_\mathbf{g}^{-\frac{1}{2}} \mathbf{K}_\mathbf{g} \mathbf{K}_\mathbf{g}^{-\frac{1}{2}} = \mathbf{I}\,, \tag{8.68}$$

where we have used the definition of the square-root matrix from (A.117) and the fact that covariance matrices are Hermitian, so that $\mathbf{K}_\mathbf{g}^{-\frac{1}{2}}$ is its own adjoint.

Thus the transformation (8.67) always results in a random vector $\mathbf{y}$ such that

$$\langle y_n y_m \rangle = \delta_{nm}\,. \tag{8.69}$$

By analogy to white noise (a topic discussed further in Sec. 8.2.6), this transformation is referred to as *whitening*; it is also called *prewhitening* when it precedes other signal processing. As we shall see in Chap. 13, prewhitening plays a key role in signal-detection theory.

Simultaneous diagonalization We have shown that a Hermitian matrix can always be diagonalized by a unitary transformation. It can be shown that two different Hermitian matrices can be diagonalized by the *same* unitary transformation if and only if they commute. If the different Hermitian matrices do not commute, they can be simultaneously diagonalized by a linear transformation, but the transformation matrix will not be unitary (Fukunaga, 1990). Details of the procedure were given in Sec. 1.4.6.

8.2 RANDOM PROCESSES

8.2.1 Definitions and basic concepts

We now generalize the concept of a random variable further by assigning to every experimental outcome ζ a spatial or temporal function, real or complex, according to some rule (Wentzell, 1981). In the spatial case the function will be denoted $f(\mathbf{r}, \zeta)$, where $\mathbf{r}$ is a position vector, and in the temporal case it will be denoted by $f(t, \zeta)$, where t is the time. We now have a family of functions referred to as a stochastic or random process. The words stochastic and random will be used interchangeably here. If the spatial (or temporal) variable $\mathbf{r}$ (or t) is a continuous one, the family is referred to as a continuous stochastic process; if the variables are taken as discrete, for example by sampling in space or time, the family is referred to as a discrete stochastic process, or a random sequence. A random process or sequence is said to be continuous-valued or discrete-valued according to whether the underlying random variables are continuous- or discrete-valued.

A spatial random process is a function of two variables, $\mathbf{r}$ and ζ. Depending on the context, $f(\mathbf{r}, \zeta)$ can refer to (Papoulis, 1965; Middleton, 1960):

1. The family of spatial functions, referred to as the ensemble; in this case, $\mathbf{r}$ and ζ are variables.

2. A single realization or sample of the spatial functions; in this case, $\mathbf{r}$ is variable and ζ is fixed.

3. The random variable at a single point; in this case, $\mathbf{r}$ is fixed and ζ is variable.

4. A single number; in this case, $\mathbf{r}$ is fixed and ζ is fixed.

The intended interpretation will usually be clear from the context.

Some notational issues require attention here. First, it is frequently cumbersome to carry along the index ζ, so we shall usually refer to the random process simply as $f(\mathbf{r})$. Second, we usually make no notational distinction between the random process *per se*, understood as the ensemble of all possible functions $f(\mathbf{r})$ (interpretation 1), and a specific realization or sample (interpretation 2). This practice is in accord with our conventions on random variables as set out in Sec. C.2.1 of App. C. Occasionally a specific realization will have to be designated, and in those cases we shall either reinstate ζ or use primes, subscripts or other typographical devices. Finally, the variable $\mathbf{r}$ will be understood to be a general q-dimensional position vector unless otherwise stated.

Square-integrable random processes We shall say that a random process lies in some Hilbert space if all sample functions [*i.e.*, $f(\mathbf{r}, \zeta)$ as a function of $\mathbf{r}$ for all ζ] lie in that space. In particular, we shall often be concerned with random processes in $\mathbb{L}_2(\mathbb{R}^q)$ where each sample function is square-integrable.

In many physical problems, $|f(\mathbf{r})|^2$ can be interpreted as an energy density (energy per unit area or volume). For example, that interpretation works when $f(\mathbf{r})$ is an electric field or the amplitude of an acoustic wave. In those cases, the integral of $|f(\mathbf{r})|^2$ is the total energy, and a square-integrable function is one with finite energy. This terminology is used more broadly, and any square-integrable function can be called a finite-energy function without regard to interpretation as a physical energy.

Finite-power random processes For temporal random processes, however, the assumption of finite energy is frequently not warranted. Consider, for example, the thermal noise produced by a resistor. The duration of this noise is completely indefinite. So long as the resistor exists, there will be a fluctuating voltage across it. In this example, ζ designates a particular resistor and $f(t, \zeta)$ is the noise voltage, and there is no reason to assume that the integral of $|f(t, \zeta)|^2$ over $-\infty < t < \infty$ is finite. We could get around this problem by imposing some artificial boundary conditions, *e.g.*, the resistor is manufactured at $t = -T$ and destroyed at $t = T$, but we are not really interested in when the resistor was manufactured.

A more natural approach is just to give up on the restriction to finite energy. The noise voltage across a resistor has finite *power* (energy per unit time). Mathematically, we can state this condition for the random process $f(t, \zeta)$ as

$$0 < \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt\, |f(t, \zeta)|^2 < \infty \qquad \text{for all } \zeta\,. \tag{8.70}$$

A random process for which this condition is satisfied will be called a *finite-power random process*. Note that a finite-energy (or $\mathbb{L}_2$) random process cannot simultaneously be a finite-power one because of the left-hand inequality. If the function is in $\mathbb{L}_2$, then the integral is finite as $T \to \infty$, but the factor of $1/2T$ drives the product to zero. It is only when the integral is asymptotically linear in T that (8.70) is satisfied. As we shall see, finite-energy and finite-power random processes require rather different mathematical treatments.

Generalized random processes We shall also have occasion to use random processes constructed with delta functions or other generalized functions. Such constructs are mathematically very convenient, even though no physical process is exactly described by them. We shall refer to random processes where the sample functions are generalized functions as (not surprisingly) *generalized random processes* (Kanwal, 1983). These processes are not in $\mathbb{L}_2$ but instead define a space of tempered distributions (see Chap. 2). If the generalized function in question is a delta function, the generalized random process has neither finite energy nor finite power.

8.2.2 Averages of random processes

We consider here a scalar random process $f(\mathbf{r})$ that is a function of position vector $\mathbf{r}$. The generalization to a vector random process is straightforward using multivariate PDFs. The random process may in principle be either continuous- or discrete-valued, but we shall illustrate the concepts with continuous-valued random processes. The discrete-valued case proceeds via a parallel approach but with sums over discrete values replacing integrals over continuous values.

For fixed $\mathbf{r}$, $f(\mathbf{r})$ is simply a random variable (interpretation 3), and its expectation is defined just as for any other random variable. As before, we use the notations $\mathrm{E}\{\cdot\}$, $\langle\cdot\rangle$ and overbar interchangeably to indicate an expectation, and we can write

$$\mathrm{E}\{f(\mathbf{r})\} = \langle f(\mathbf{r})\rangle = \overline{f}(\mathbf{r}) = \int_{-\infty}^{\infty} df(\mathbf{r})\; f(\mathbf{r})\,\mathrm{pr}[f(\mathbf{r})]\,. \tag{8.71}$$

Computation of this expectation requires only the univariate PDF $\mathrm{pr}[f(\mathbf{r})]$. Note carefully that the integral is over $f(\mathbf{r})$, not $\mathbf{r}$, so $\mathrm{E}\{f(\mathbf{r})\}$ can be (and usually will be) a function of $\mathbf{r}$.

Moments and variance Moments of $f(\mathbf{r})$ are defined easily. For example, the j^{th} moment is given by [*cf.* (C.38)]

$$\left\langle [f(\mathbf{r})]^j \right\rangle = \int_{-\infty}^{\infty} df(\mathbf{r})\; [f(\mathbf{r})]^j\,\mathrm{pr}[f(\mathbf{r})]\,. \tag{8.72}$$

The resultant, $\left\langle [f(\mathbf{r})]^j \right\rangle$, can still be a function of $\mathbf{r}$; again, the integral is over $f(\mathbf{r})$, not over $\mathbf{r}$.

Having defined moments, we can also define the variance of a random process. In the general complex case, the variance is given by

$$\begin{aligned}\mathrm{Var}\{f(\mathbf{r})\} &= \mathrm{E}\left\{|f(\mathbf{r})| - |\mathrm{E}\{f(\mathbf{r})\}|^2\right\} = \mathrm{E}\left\{|f(\mathbf{r})|^2\right\} - |\mathrm{E}\{f(\mathbf{r})\}|^2 \\ &= \int_{-\infty}^{\infty} df(\mathbf{r})\; |f(\mathbf{r})|^2\,\mathrm{pr}[f(\mathbf{r})] - \left|\int_{-\infty}^{\infty} df(\mathbf{r})\; f(\mathbf{r})\,\mathrm{pr}[f(\mathbf{r})]\right|^2 .\end{aligned} \tag{8.73}$$

Note that this definition works equally for finite-energy and finite-power processes. It is possible for a random process to have a finite variance at all points, yet not be square-integrable.

Multiple-point expectations We are often interested in *two-point expectations* or joint second moments of the form $\mathrm{E}\{f(\mathbf{r}_1)f(\mathbf{r}_2)\}$. The usual definitions for joint expectations stand us in good stead here, and we can write

$$\mathrm{E}\{f(\mathbf{r}_1)f(\mathbf{r}_2)\} = \int_{-\infty}^{\infty} df(\mathbf{r}_1) \int_{-\infty}^{\infty} df(\mathbf{r}_2)\;\; f(\mathbf{r}_1)\,f(\mathbf{r}_2)\,\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]\,. \tag{8.74}$$

Here, $f(\mathbf{r}_1)$ and $f(\mathbf{r}_2)$ must be regarded as two *distinct* random variables and $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]$ is their joint density. Only in very special circumstances will it be possible to write $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]$ as $\mathrm{pr}[f(\mathbf{r}_1)]\,\mathrm{pr}[f(\mathbf{r}_2)]$.

A general two-point moment is defined by

$$\mathrm{E}\big\{[f(\mathbf{r}_1)]^m\,[f(\mathbf{r}_2)]^n\big\} = \int_{-\infty}^{\infty} df(\mathbf{r}_1) \int_{-\infty}^{\infty} df(\mathbf{r}_2)\;\, [f(\mathbf{r}_1)]^m\,[f(\mathbf{r}_2)]^n\,\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]\,. \tag{8.75}$$

Moments involving more points are defined similarly. Any moment involving the K points $\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_K$ can be computed if $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2), ..., f(\mathbf{r}_K)]$ is known. If this K-fold joint density is known for all values of each of the $\mathbf{r}_k$, the process is said to be *fully characterized* to order K (Snyder and Miller, 1991).

Density of the process Expressing N-fold joint densities using notation of the form $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2), ..., f(\mathbf{r}_N)]$ is cumbersome at best and quite inadequate when we want to define expectations of general functionals $\Phi\{f(\mathbf{r})\}$, which can depend on $f(\mathbf{r})$ at all points $\mathbf{r}$. We now introduce an alternative approach, which works at least for finite-energy random processes (or vectors in $\mathbb{L}_2$). Our objective is to give meaning to an expression like $\mathrm{pr}(\mathbf{f})$, where $\mathbf{f}$ is the Hilbert-space vector corresponding to $f(\mathbf{r})$. We saw in Chap. 1 that $\mathbb{L}_2$ is a *separable* Hilbert space, which means simply that it is spanned by a denumerably infinite set of basis functions. Each sample function of a random process in $\mathbb{L}_2(\mathbb{R}^q)$ can be written as

$$f(\mathbf{r}, \zeta) = \sum_{j=1}^{\infty} \alpha_j(\zeta)\, \psi_j(\mathbf{r}) \,, \tag{8.76}$$

where the set $\{\psi_j(\mathbf{r})\}$ is any orthonormal basis for the space. We can also express this same concept as

$$f(\mathbf{r}) = \lim_{J\to\infty} \sum_{j=1}^{J} \alpha_j \psi_j(\mathbf{r}) \,. \tag{8.77}$$

We have dropped the index ζ with the understanding that the equation holds for any $f(\mathbf{r}, \zeta)$ so long as the corresponding expansion coefficients $\alpha_j(\zeta)$ are used on the right. The convergence of (8.77) is in the sense of $\mathbb{L}_2(\mathbb{R}^q)$ (see Sec. 3.2.2); if we use the truncated series in place of the original function $f(\mathbf{r})$, the $\mathbb{L}_2$ norm of the error converges to zero as $J \to \infty$. The expansion of the sample function $f(\mathbf{r})$ given in (8.77) is exactly the same form that was used in (7.8) to represent a deterministic object.

Expansion (8.77) provides a convenient way of defining averages involving random processes. Each coefficient α_j is a random variable, and the set of them $\{\alpha_j, j = 1, ..., J\}$ can be regarded as a random vector $\boldsymbol{\alpha}_J$ with J components. In the limit $J \to \infty$, the vector $\boldsymbol{\alpha}_J$ completely defines $f(\mathbf{r})$, and averaging over $f(\mathbf{r})$ is equivalent to averaging over all components of $\boldsymbol{\alpha}$. For finite J, the requisite density can be written as $\mathrm{pr}(\boldsymbol{\alpha}_J)$ or $\mathrm{pr}(\alpha_1, \alpha_2, ..., \alpha_J)$. In the limit,

$$\mathrm{pr}(\boldsymbol{\alpha}) = \lim_{J\to\infty} \mathrm{pr}(\boldsymbol{\alpha}_J) \,, \tag{8.78}$$

and this density is operationally equivalent to $\mathrm{pr}(\mathbf{f})$.

When $f(\mathbf{r})$ is approximated by the truncated series, any functional $\Phi\{f(\mathbf{r})\}$ is also a function of $\boldsymbol{\alpha}_J$; call it $\Phi_J(\boldsymbol{\alpha}_J)$. If the functional is continuous, in the sense defined in Sec. 1.3.2, then the limit of the functional is the functional of the limit, and we have

$$\Phi\{f(\mathbf{r})\} = \lim_{J\to\infty} \Phi_J(\boldsymbol{\alpha}_J) \,. \tag{8.79}$$

Moreover, expectation is also a continuous functional, so we can write

$$\mathrm{E}\{\Phi[f(\mathbf{r})]\} = \lim_{J\to\infty} \mathrm{E}\{\Phi_J(\boldsymbol{\alpha}_J)\} = \lim_{J\to\infty} \int_\infty d^J\alpha\; \Phi_J(\boldsymbol{\alpha}_J)\, \mathrm{pr}(\boldsymbol{\alpha}_J) \,. \tag{8.80}$$

For notational convenience, we write this expectation as

$$\mathrm{E}\{\Phi[f(\mathbf{r})]\} = \int_{\mathbb{L}_2} d\mathbf{f}\ \Phi[f(\mathbf{r})]\,\mathrm{pr}(\mathbf{f})\,. \tag{8.81}$$

Here $\mathbf{f}$ is $f(\mathbf{r})$ regarded as a vector in the Hilbert space, and the integral is really a denumerably infinite multiple integral[2] over all basis functions in the space; in other words, (8.81) must be realized operationally by (8.80).

Example: Linear functionals To clarify how (8.81) works in practice, consider a linear functional that depends on $f(\mathbf{r})$ at K points:

$$\Phi\{f(\mathbf{r}_1), ..., f(\mathbf{r}_K)\} \equiv \sum_{k=1}^{K} \beta_k f(\mathbf{r}_k) = \lim_{J\to\infty} \sum_{k=1}^{K} \beta_k \sum_{j=1}^{J} \alpha_j \psi_j(\mathbf{r}_k)\,. \tag{8.82}$$

The random variables here are the coefficients $\{\alpha_j\}$. Using (8.80) and invoking the linearity of the expectation operator, we find

$$\mathrm{E}\{\Phi[f(\mathbf{r}_1), ..., f(\mathbf{r}_K)]\} = \lim_{J\to\infty} \sum_{k=1}^{K} \beta_k \sum_{j=1}^{J} \psi_j(\mathbf{r}_k) \int_{\infty} d^J\alpha\ \alpha_j\, \mathrm{pr}(\boldsymbol{\alpha}_J)\,. \tag{8.83}$$

In the J-fold multiple integral, we can immediately integrate out all of the variables except α_j. By (C.75), the result is the marginal density on α_j, so

$$\begin{aligned}\mathrm{E}\{\Phi[f(\mathbf{r}_1), ..., f(\mathbf{r}_K)]\} &= \lim_{J\to\infty} \sum_{k=1}^{K} \beta_k \sum_{j=1}^{J} \psi_j(\mathbf{r}_k) \int_{-\infty}^{\infty} d\alpha_j\ \alpha_j\, \mathrm{pr}(\alpha_j) \\ &= \lim_{J\to\infty} \sum_{k=1}^{K} \beta_k \sum_{j=1}^{J} \psi_j(\mathbf{r}_k)\, \mathrm{E}\{\alpha_j\}\,. \end{aligned} \tag{8.84}$$

Thus, for a linear functional of the form (8.82), and by extension any linear functional,

$$\langle \Phi\{\mathbf{f}\}\rangle = \Phi\{\langle \mathbf{f}\rangle\}\,. \tag{8.85}$$

Integrals of random processes An integral of a random process $f(x)$, sometimes called a *stochastic integral*, is another random process, the realizations of which are obtained by integrating corresponding realizations of $f(x)$. For example, the statement

$$g(x) = \int_{-\infty}^{\infty} dx'\, f(x')\, h(x, x') \tag{8.86}$$

means that

$$g(x, \zeta) = \int_{-\infty}^{\infty} dx'\, f(x', \zeta)\, h(x, x') \tag{8.87}$$

---

[2] We have customarily denoted volume elements by italics rather than boldface, *e.g.*, d^3r rather than $d\mathbf{r}$, on the theory that volume elements are scalars. To preserve a distinction between integrals over Euclidean spaces and ones over Hilbert spaces, however, we use $d\mathbf{f}$ (along with the subscript $\mathbb{L}_2$) to indicate a multiple integral with an infinite number of dimensions.

for all ζ and some fixed kernel $h(x, x')$. A similar definition holds for derivatives of a random process.

Since $g(x)$ is a functional of $f(x')$, its average at any fixed x can be computed by (8.81) as

$$\langle g(x) \rangle = \int_{\mathbb{L}_2} d\mathbf{f}\; g(x)\,\mathrm{pr}(\mathbf{f}) = \int_{\mathbb{L}_2} d\mathbf{f} \int_{-\infty}^{\infty} dx'\, f(x')\, h(x, x')\,\mathrm{pr}(\mathbf{f})\,. \tag{8.88}$$

It is often useful to interchange the order of these two integrals, but most books gloss over issues of the validity of this step. Middleton (1960) puts it thus: "The condition on the interchangeability of integration and expectation is, *of course*, the existence of the resulting integral" (emphasis added).

When the interchange can be justified, (8.88) becomes

$$\langle g(x) \rangle = \int_{-\infty}^{\infty} dx'\, h(x, x') \int_{\mathbb{L}_2} d\mathbf{f}\; f(x')\,\mathrm{pr}(\mathbf{f}) = \int_{-\infty}^{\infty} dx'\, h(x, x')\,\langle f(x') \rangle\;. \tag{8.89}$$

In other words, the average of a linear integral transform of a random process is the same linear transform of the average of the random process (but only under conditions that we haven't yet stated clearly).

When is the interchange legal? The classical theorem that states when interchange of the order of two integrals is allowed is Fubini's theorem (Lang, 1993). In essence, this theorem tells us that

$$\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\; k(u, v) = \int_{-\infty}^{\infty} du \left[\int_{-\infty}^{\infty} dv\; k(u, v)\right] = \int_{-\infty}^{\infty} dv \left[\int_{-\infty}^{\infty} du\; k(u, v)\right] \tag{8.90}$$

provided $|k(u, v)|$ is integrable over the product space, here the u-v plane.

There are two difficulties in directly applying Fubini to (8.88). First, we often want to assume that the integrand is in $\mathbb{L}_2$ rather than in $\mathbb{L}_1$, and we know from Sec. 3.3.2 that a function in $\mathbb{L}_2$ need not be in $\mathbb{L}_1$ (the prime example being $\operatorname{sinc} x$). One way around this problem is to consider only random processes where all sample functions are absolutely integrable as well as square-integrable. Another way is to consider a finite interval, say $-\frac{1}{2}X < x \leq \frac{1}{2}X$. This allows use of Fubini with $\mathbb{L}_2$ functions since $\mathbb{L}_2(-\frac{1}{2}X, \frac{1}{2}X)$ is a subspace of $\mathbb{L}_1(-\frac{1}{2}X, \frac{1}{2}X)$.

The second difficulty is that Fubini's theorem can be extended to higher-dimensional multiple integrals, but (8.88) in its most general form requires an infinite nested set of integrals. Fubini's theorem can still be used in this case, but it must be justified with advanced measure-theoretic arguments (Lipster and Shiryayev, 1977). A more elementary argument can be given by using the theory of distributions.

Retreat to distributions Much of the discussion above has hinged on the assumption that the random process lies in a separable Hilbert space. For finite-power processes, we do not have this luxury, and even with $\mathbb{L}_2$ processes, we ran into some problems justifying the interchange of integration and expectation. The solution to these difficulties is the theory of distributions[3] as outlined in Chap. 2. The thing we have

---

[3] At least three distinctly different meanings attach to the word *distribution* in connection with random processes. A *probability distribution* is, loosely speaking, any probability law, such as the

going for us is that sample functions of a random process may be badly behaved but kernel functions in integral transforms like (8.86) are usually good functions.

Let $t(x)$ denote a good function and $f(x,\zeta)$ be a sample function of a random process. This random process defines a distribution,

$$\Phi_f\left\{t(x)\right\} = \int_{-\infty}^{\infty} dx\; t(x)\, f(x,\zeta) \equiv \phi(\zeta)\,. \tag{8.91}$$

Note that $\phi(\zeta)$ is a random variable. It is proved by Kanwal (1983) that this random variable has finite variance if $f(x,\zeta)$ is continuous (in the sense that $f(x+\epsilon,\zeta) \to f(x,\zeta)$ in the limit that $\epsilon \to 0$) and has finite variance at all x. With these mild restrictions, any random process defines a distribution mapping good functions to finite-variance random variables.

By the Schwarz inequality, the finite variance of $\phi(\zeta)$ implies that $\phi(\zeta)$ has finite mean. The expectation $\mathrm{E}\{\phi(\zeta)\}$ is defined conventionally by

$$\mathrm{E}\{\phi(\zeta)\} = \mathrm{E}\{\Phi_f[t(x)]\} = \int_{-\infty}^{\infty} d\phi\; \phi\, \mathrm{pr}(\phi)\,. \tag{8.92}$$

But this is just a linear combination of distributions, which by (2.15) is another distribution. Thus

$$\mathrm{E}\{\Phi_f[t(x)]\} = \Phi_{Ef}\{t(x)\} = \int_{-\infty}^{\infty} dx\; t(x)\, \mathrm{E}\{f(x,\zeta)\}\,, \tag{8.93}$$

where $\Phi_{Ef}\{t(x)\}$ is a distribution defined by using $\mathrm{E}\{f(x,\zeta)\}$ as the generalized function. Equation (8.93) is just what one would obtain by interchanging the expectation operation and the integration over x.

Thus the issue of interchangeability is resolved once we have established that the random process defines a distribution (in the Schwartz sense), and Kanwal did this for us with mild restrictions.

8.2.3 Characteristic functionals

Characteristic functions for scalar random variables were introduced in App. C and extended to random vectors in Sec. 8.1.4. Now we shall extend the concept further to random processes. In a formal sense, the extension is straightforward; all we have to do is to pay attention to the dimensionality of the vectors involved.

As defined in (8.26), the characteristic function of an MD random vector is a function of an MD frequency vector $\boldsymbol{\xi}$. In the case of a random process, each sample function corresponds to a vector $\mathbf{f}$ in an infinite-dimensional Hilbert space, so the frequency vector $\boldsymbol{\xi}$ in (8.26) must be replaced by an infinite-dimensional vector $\mathbf{s}$ in the same Hilbert space as $\mathbf{f}$. That means that $\mathbf{s}$ describes a function $s(\mathbf{r})$, so the characteristic function becomes a characteristic *functional* $\Psi_{\mathbf{f}}\{s(\mathbf{r})\}$ or $\Psi_{\mathbf{f}}(\mathbf{s})$ for short. It is defined by

$$\Psi_{\mathbf{f}}(\mathbf{s}) = \langle \exp[-2\pi i(\mathbf{s},\mathbf{f})]\rangle\,, \tag{8.94}$$

Poisson distribution. The *distribution function* refers specifically to the cumulative probability distribution function defined in Sec. C.2.3. In the present context the word is used in the Schwartz sense defined in Chap. 2.

where $(\mathbf{s}, \mathbf{f})$ is the usual $\mathbb{L}_2$ scalar product. Note that we use $\Psi(\,\cdot\,)$ for characteristic functional and $\psi(\,\cdot\,)$ for characteristic function.

The characteristic functional of a random process can be related to the characteristic *function* of any random vector derived from the random process by a linear operator; the calculation is a simple generalization of one performed in Sec. 8.1.5. For example, if $\mathbf{g} = \boldsymbol{\mathcal{H}}\mathbf{f}$, where $\boldsymbol{\mathcal{H}}$ is a continuous-to-discrete (CD) operator as discussed in Secs. 1.2.4 and 7.3, then (8.26) becomes

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \langle \exp[-2\pi i(\boldsymbol{\xi}, \boldsymbol{\mathcal{H}}\mathbf{f})] \rangle = \langle \exp[-2\pi i(\boldsymbol{\mathcal{H}}^\dagger \boldsymbol{\xi}, \mathbf{f})] \rangle\,, \tag{8.95}$$

where the second step follows from the definition of the adjoint, (1.39). Comparison of (8.94) and (8.95) shows that

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \Psi_{\mathbf{f}}(\boldsymbol{\mathcal{H}}^\dagger \boldsymbol{\xi})\,, \tag{8.96}$$

which is the generalization of (8.43) to random processes.

Thus, if we know the characteristic functional for $\mathbf{f}$, we immediately have the characteristic function for $\boldsymbol{\mathcal{H}}\mathbf{f}$. We shall exploit this relation in Sec. 8.3.5 when we discuss normal random processes.

8.2.4 Correlation analysis

The autocorrelation function $R(\mathbf{r}_1, \mathbf{r}_2)$ of a random process $f(\mathbf{r})$ is defined by

$$R(\mathbf{r}_1, \mathbf{r}_2) = \langle f(\mathbf{r}_1)\, f^*(\mathbf{r}_2) \rangle\;, \tag{8.97}$$

which is the two-point expectation defined in (8.74), with the minor modification of the complex conjugate on the second factor [irrelevant if $f(\mathbf{r})$ is real].

The autocovariance function $K(\mathbf{r}_1, \mathbf{r}_2)$ is defined by

$$\begin{aligned} K(\mathbf{r}_1, \mathbf{r}_2) &= \langle [f(\mathbf{r}_1) - \langle f(\mathbf{r}_1) \rangle]\, [f^*(\mathbf{r}_2) - \langle f^*(\mathbf{r}_2) \rangle] \rangle \\ &= R(\mathbf{r}_1, \mathbf{r}_2) - \overline{f}(\mathbf{r}_1)\, \overline{f}^*(\mathbf{r}_2)\,. \end{aligned} \tag{8.98}$$

The autocovariance function is thus the two-point moment that is the generalization of the variance; it reduces to the variance when $\mathbf{r}_2 = \mathbf{r}_1 = \mathbf{r}$, *i.e.*,

$$K(\mathbf{r}, \mathbf{r}) = R(\mathbf{r}, \mathbf{r}) - |\overline{f}(\mathbf{r})|^2 = \mathrm{Var}\{f(\mathbf{r})\} \tag{8.99}$$

from (8.73).

When two or more random processes occur in the same problem, their autocorrelation and autocovariance functions will be distinguished with subscripts, *e.g.*, $R_f(\mathbf{r}_1, \mathbf{r}_2)$. It is frequently convenient to define zero-mean random processes such as

$$\Delta f(\mathbf{r}) \equiv f(\mathbf{r}) - \overline{f}(\mathbf{r})\,. \tag{8.100}$$

With this definition, $\langle \Delta f(\mathbf{r}) \rangle = 0$ and

$$R_{\Delta f}(\mathbf{r}_1, \mathbf{r}_2) = K_f(\mathbf{r}_1, \mathbf{r}_2)\,. \tag{8.101}$$

The autocorrelation and autocovariance functions play a fundamental role in the theory of random processes since they specify how far apart two points must be for their fluctuations to be uncorrelated. If $K_f(\mathbf{r}_1, \mathbf{r}_2)$ is zero, the random variables

$f(\mathbf{r}_1)$ and $f(\mathbf{r}_2)$ do not covary; colloquially, they are said to be uncorrelated, though in fact the autocorrelation function $R_f(\mathbf{r}_1, \mathbf{r}_2)$ may be nonzero because of the mean values.

Cross-correlation and cross-covariance functions can also be defined. Consider two functions $f(\mathbf{r})$ and $g(\mathbf{r}')$, where $\mathbf{r}$ and $\mathbf{r}'$ are not necessarily in the same space. The cross-correlation or mutual correlation function is defined by

$$R_{fg}(\mathbf{r}, \mathbf{r}') = \langle f(\mathbf{r})\, g^*(\mathbf{r}') \rangle \,. \tag{8.102}$$

Similarly, the cross-covariance function is

$$K_{fg}(\mathbf{r}, \mathbf{r}') = \langle [f(\mathbf{r}) - \langle f(\mathbf{r}) \rangle]\, [g^*(\mathbf{r}') - \langle g^*(\mathbf{r}') \rangle] \rangle = R_{fg}(\mathbf{r}, \mathbf{r}') - \overline{f}(\mathbf{r})\, \overline{g}^*(\mathbf{r}') \,. \tag{8.103}$$

Two random processes $f(\mathbf{r})$ and $g(\mathbf{r}')$ are said to be uncorrelated if $R_{fg}(\mathbf{r}, \mathbf{r}') = \overline{f}(\mathbf{r})\, \overline{g}(\mathbf{r}')$ for all $\mathbf{r}$ and $\mathbf{r}'$. They are orthogonal if, for all $\mathbf{r}$ and $\mathbf{r}'$, $R_{fg}(\mathbf{r}, \mathbf{r}') = 0$.

Properties of the autocorrelation function From the definition (8.98), we obtain the symmetry property

$$R(\mathbf{r}_1, \mathbf{r}_2) = R^*(\mathbf{r}_2, \mathbf{r}_1) \,. \tag{8.104}$$

In particular, for $\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r}$, (8.104) shows that $R(\mathbf{r}, \mathbf{r})$ or $\mathrm{Var}\{f(\mathbf{r})\}$ is real.

It follows from the Schwarz inequality that

$$|R(\mathbf{r}_1, \mathbf{r}_2)|^2 \leq R(\mathbf{r}_1, \mathbf{r}_1)\, R(\mathbf{r}_2, \mathbf{r}_2) \,. \tag{8.105}$$

It can also be shown (Mandel and Wolf, 1995) that $R(\mathbf{r}_1, \mathbf{r}_2)$ is positive-semidefinite, meaning that [*cf.* (8.22)]

$$\int_\infty d^q r_1\, w^*(\mathbf{r}_1) \int_\infty d^q r_2\ R(\mathbf{r}_1, \mathbf{r}_2)\, w(\mathbf{r}_2) \geq 0 \,, \tag{8.106}$$

for all functions $w(\mathbf{r})$. We shall exploit this property in Sec. 8.2.7 when we discuss the Karhunen-Loève expansion of random processes.

Another way to think about $R(\mathbf{r}_1, \mathbf{r}_2)$ is that it is the kernel of an integral operator $\boldsymbol{\mathcal{R}}$. With this view, the inner integral of (8.106) is recognized as the function $[\boldsymbol{\mathcal{R}}\mathbf{w}](\mathbf{r}_1)$, and the double integral is the scalar $\mathbf{w}^\dagger \boldsymbol{\mathcal{R}} \mathbf{w}$. An autocovariance operator $\boldsymbol{\mathcal{K}}$ can be defined similarly, with the autocovariance function as the kernel.

Temporal stationarity Temporal random processes often have a statistical character that is independent of time, even though any individual realization is a randomly fluctuating function of time. An example is a steady beam of white light, where the electric field fluctuates rapidly, yet there is no preferred origin in time as far as the statistics are concerned. Such processes are said to be *stationary*. Glauber (1965) has phrased it this way: "The term 'stationary' does not mean that nothing is happening. On the contrary, the field is ordinarily oscillating quite rapidly. It means that our knowledge of the field does not change in time."

A temporal random process $f(t)$ is said to be stationary in the strict sense if, for any K, its K-point PDF $\mathrm{pr}[f(t_1), \cdots, f(t_K)]$ is such that

$$\mathrm{pr}[f(t_1), \cdots, f(t_K)] = \mathrm{pr}[f(t_1 + \tau), \cdots, f(t_K + \tau)] \tag{8.107}$$

for any τ. In particular, this requires that the single-point density function be independent of time,

$$\mathrm{pr}[f(t)] = \mathrm{pr}[f(t + \tau)] \,, \tag{8.108}$$

and therefore the mean of the random process is also independent of time,

$$\langle f(t)\rangle = \langle f(t+\tau)\rangle \ . \tag{8.109}$$

Similarly, the two-point density function must be independent of time,

$$\mathrm{pr}[f(t_1), f(t_2)] = \mathrm{pr}[f(t_1+\tau), f(t_2+\tau)]\,, \tag{8.110}$$

and so the autocorrelation function $R(t_1, t_2)$, is also independent of time,

$$R(t_1, t_2) = \langle f(t_1)\, f^*(t_2)\rangle = \langle f(t_1+\tau)\, f^*(t_2+\tau)\rangle \ . \tag{8.111}$$

The only way (8.111) can be satisfied for all t_1 and t_2 is if $R(t_1, t_2)$ is really a function of only $t_1 - t_2$. We shall denote this function by $R(t_1 - t_2)$, but the reader is cautioned that $R(t_1 - t_2)$ is not the same function as $R(t_1, t_2)$; it could not be since the latter has two arguments and the former has only one. With this notation, we have (for a stationary random process),

$$R(t_1, t_2) = R(t_1 - t_2) = R(\Delta t)\,, \tag{8.112}$$

where $\Delta t \equiv t_1 - t_2$. The shift Δt is frequently called the *lag* of the autocorrelation function.

Continuing on in this way, we see that strict stationarity requires that all K-point moments of the process be independent of absolute time. A process is said to be stationary to order M if (8.107) is true only for $K \leq M$.

A process is said to be weakly stationary, or stationary in the wide sense, if its expected value does not depend on absolute time t and its autocorrelation depends only on Δt:

$$\langle f(t)\rangle = const\,, \qquad \langle f(t+\Delta t)\, f^*(t)\rangle = R(\Delta t)\,. \tag{8.113}$$

If a process is stationary to second order, then it is wide-sense stationary; however, a wide-sense stationary process is not necessarily stationary to second order because the former involves only the first two moments while the latter involves the entire PDF. One case where we can make a more definitive statement is with normal or Gaussian random processes, to be discussed in Sec. 8.3.5. If a process is normal and stationary in the wide sense, then it is also stationary in the strict sense since the statistical description of a normal process is completely specified once its mean and autocorrelation are specified.

Stationarity is closely connected to the concept of a finite-power random process, introduced in Sec. 8.2.1, but the distinctions should not be overlooked. The finite-power designation applies to individual sample functions of the random process, while stationarity applies to averages. A stationary random process might not have finite power, since it is conceivable (though pathological) that an individual realization could diverge but the average not. Of more practical importance, a process can have finite power yet not be stationary; examples of this situation are discussed below. On the other hand, a nontrivial stationary temporal random process certainly cannot have finite energy.

Properties of the stationary autocorrelation function The general properties of autocorrelations given above specialize in the stationary case as follows: The symmetry property of (8.104) becomes

$$R(\Delta t) = R^*(-\Delta t)\,. \tag{8.114}$$

In particular, for $\Delta t = 0$, (8.114) shows that $R(0)$ is real.

The Schwarz inequality shows that

$$|R(\Delta t)| \leq R(0)\,. \tag{8.115}$$

The condition that $R(\Delta t)$ is positive-semidefinite now means that

$$\int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} dt'\, w^*(t)\, R(t-t')\, w(t') \geq 0\,, \tag{8.116}$$

for all functions $w(t)$.

Spatial stationarity The spatial counterpart of the wide-sense stationarity condition (8.112) is

$$R(\mathbf{r}_1, \mathbf{r}_2) = R(\mathbf{r}_1 - \mathbf{r}_2) = R(\Delta\mathbf{r})\,, \tag{8.117}$$

where $\Delta\mathbf{r} \equiv \mathbf{r}_1 - \mathbf{r}_2$.

This condition cannot be exactly satisfied[4] by spatial processes representing real objects or images since they have finite support, but it might be a useful mathematical description within a certain boundary. That is, we might be able to assume that $R(\mathbf{r}_1, \mathbf{r}_2) = R(\Delta\mathbf{r})$ provided $\mathbf{r}_1$ and $\mathbf{r}_2$ lie inside the borders of an image. An example would be a piece of x-ray film with a uniform exposure, where the only deviation from stationarity comes from the finite size of the film.

If $f(\mathbf{r})$ vanishes outside the boundary, this kind of stationarity is expressed mathematically by

$$R(\mathbf{r}_1, \mathbf{r}_2) = R(\Delta\mathbf{r})\, W(\mathbf{r}_1)\, W(\mathbf{r}_2)\,, \tag{8.118}$$

where $W(\mathbf{r})$ is a window function that is unity for $\mathbf{r}$ inside the boundary, zero outside.

Quasistationarity In optics and imaging we often encounter spatial random processes whose autocorrelation function can be approximated as a product of two factors — a slowly varying contribution due to slow variations in overall intensity and a short-range function describing correlation between neighboring points. As a simple example, consider a ground glass illuminated nonuniformly with a laser beam. If the statistical character of the ground glass is the same at all points, then we can describe the complex amplitude (see Chap. 9) of the wave emerging from the ground glass by a spatial autocorrelation function of the form,

$$R(\mathbf{r}_1, \mathbf{r}_2) = a(\Delta\mathbf{r})\, b(\mathbf{r}_0)\,, \tag{8.119}$$

where

$$\mathbf{r}_0 = \tfrac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2)\,, \qquad \Delta\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2\,. \tag{8.120}$$

We shall refer to $\mathbf{r}_0$ as the *center coordinate* (analogous to center of mass) and $\Delta\mathbf{r}$ as the *relative coordinate* or *difference coordinate*. Since the transformation from $(\mathbf{r}_1, \mathbf{r}_2)$ to $(\mathbf{r}_0, \Delta\mathbf{r})$ is unique and invertible (with Jacobian = unity), we always have a choice of which coordinate system to use for any function of two variables, but we

---

[4] The stationarity condition cannot be exactly satisfied by real temporal processes either. The difference is that we usually do not observe the beginning and end of a temporal process; we almost always observe the boundaries of an object or image.

won't always find that the function can be factored as in (8.119). The factorization is particularly useful if $b(\mathbf{r}_0)$ is slowly varying, in which case the random process is said to be *quasistationary.* If $b(\mathbf{r}_0)$ is a constant and the mean is also constant, the process is wide-sense stationary.

The short-range contribution, $a(\Delta\mathbf{r})$, is usually normalized to be unity at zero shift or lag ($\Delta\mathbf{r} = 0$).

Time averages and ergodicity We have seen that statistical descriptors of a random process, like the mean and autocorrelation function, are determined by averaging over the ensemble of realizations. Knowledge of the ensemble is equivalent to knowledge of the full PDF that describes the random process. However, suppose we are presented with data derived from a single realization of a temporal random process. It is natural to ask how this single data realization might be related to the statistical descriptors of the random process from which it was drawn. The answer to this question rests in the theory of *ergodicity*, a subject that traces its origins to classical statistical mechanics and the works of such luminaries as Maxwell, Boltzmann, Clausius and Gibbs (Ter Haar, 1955).

A random process is said to be *ergodic* if each realization of the process carries the same statistical information as every other realization. The practical ramification of this feature is that when a process is ergodic it becomes possible to derive statistical information about the entire ensemble based on knowledge of a single realization.

In order for a random process to be ergodic, it must first be stationary. The degree of stationarity of the process influences the degree to which the process is ergodic. For example, only wide-sense stationarity is necessary (though not sufficient) for a process to be ergodic in its mean and autocorrelation.

We now present criteria for a random process to be ergodic with respect to its mean and autocorrelation. A more complete development can be found in Papoulis (1965). Let $f(t, \zeta_0)$ denote a particular realization of a random process. Its finite-time average is then given by

$$\langle f(t, \zeta_0) \rangle_T = \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} dt \; f(t, \zeta_0) \,, \tag{8.121}$$

where $\langle \, \rangle_T$ denotes a finite-time average over period T. In general this finite-time average is itself a random variable that depends on the particular realization under consideration as well as the interval T.

The time average of the sample function $f(t, \zeta_0)$ is found by taking the limit of (8.121) as $T \to \infty$:

$$\langle f(t, \zeta_0) \rangle_\infty = \lim_{T \to \infty} \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} dt \; f(t, \zeta_0) \,. \tag{8.122}$$

The result in (8.122) is independent of time but depends in general on the realization ζ_0. Thus the notational distinction that this average refers to realization ζ_0 must be maintained.

A process is said to be *ergodic in the mean* if the time average of a single realization equals the ensemble average $\langle f(t) \rangle$. We already know that a stationary process has a mean that is independent of time. It can be shown (Papoulis, 1965) that $\langle f(t,\zeta_0)\rangle_T$ approaches this same constant as $T \to \infty$ if and only if

$$\lim_{T\to\infty} \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} d\Delta t \; R(\Delta t) = \langle f(t)\rangle^2 \,, \tag{8.123}$$

where $R(\Delta t)$ is the ensemble autocorrelation function of the stationary random process [*cf.* (8.112)]. In words, (8.123) states that ergodicity in the mean requires the time average of the autocorrelation function of $f(t)$ to be equal to the square of the ensemble mean. When this is true, the variance of the random variable that is the outcome of (8.121) approaches zero as the period T goes to infinity.

As Khinchin (1949) and others have noted, ergodicity in the mean is equivalent to the law of large numbers. In his discussion of ergodicity in statistical mechanics, Khinchin defines an ergodic process as: "On average, a system, whose evolution in time is governed by the equations of motion, remains in different parts of a given manifold of constant energy for fractions of the total time interval which are proportional to the volumes of these parts. Therefore, if we observe any physical quantity associated with a given system over a definite time interval, the arithmetic average of the results of a sufficiently large number of measurements will, as a rule, be close to the (theoretical) statistical average." He goes on to say that it is "hard to prove ergodicity in classical systems and impossible in principle to do so in quantum mechanics."

Multiple-point expectations of one realization of a temporal random process (see Sec. 8.2.2) can also be considered. For example, the finite-time autocorrelation function of one realization with itself is given by

$$R_T(\Delta t,\zeta_0) = \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} dt \; f(t+\Delta t,\zeta_0)\, f^*(t,\zeta_0) \,. \tag{8.124}$$

A random process is said to be *ergodic in autocorrelation* if $R_T(\Delta t,\zeta_0)$ approaches the ensemble quantity $R(\Delta t)$ as $T \to \infty$. We can see that the ensemble average of the sample quantity $R_T(\Delta t,\zeta_0)$ is equal to the ensemble autocorrelation function:

$$\mathrm{E}\{R_T(\Delta t,\zeta_0)\} = \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} dt \; \mathrm{E}\{f(t+\Delta t,\zeta_0)\, f^*(t,\zeta_0)\} = R(\Delta t) \,, \tag{8.125}$$

where the last step follows since $R(\Delta t)$ is independent of the integration time T. It is more difficult to demonstrate that the temporal average of $R_T(\Delta t,\zeta_0)$ approaches $R(\Delta t)$ in the limit as T becomes infinite. While a test for ergodicity of the mean requires knowledge of the ensemble mean and autocorrelation function, Papoulis demonstrates that knowledge of fourth-order moments is required to test for ergodicity of the autocorrelation function.

In general, demonstration of higher levels of ergodicity requires increasing knowledge of the density function that describes the random process. One exception, however, is the special case of the Gaussian random process. We shall see in Sec. 8.3.5 that in that case a straightforward criterion for complete ergodicity can be stated.

Ergodicity comes into play in optics when we consider the output of a detector sensing a rapidly fluctuating optical field. The period of integration in the finite-time average (8.121) is directly analogous to the detector response time. If the field fluctuates rapidly enough that fluctuations in the random process are not evident in the detector output, the random process can be said to be ergodic, and the detector can be assumed to sense an ensemble average.

We have deliberately discussed ergodicity in terms of temporal rather than spatial random processes. Remember that the first condition for ergodicity is that the random process be stationary, but as we stated earlier in this section, the physical boundaries of objects and images make spatial stationarity rarely a plausible assumption. Nevertheless, ergodicity is often assumed in the image-processing community to determine, for example, noise statistics at a single location in an image (an ensemble quantity) based on the characteristics of the fluctuations in a spatial region of that single image.

8.2.5 Spectral analysis

The Fourier transform is an important tool in the analysis of signals in general, and random signals are no exception. The Fourier transform of one sample function of a random process is defined just as for any other function, and all of the properties given in Chap. 3 are applicable. In some cases, particularly finite-power random processes, it may be necessary to consider the sample function as a generalized function and compute its Fourier transform by use of the theory of tempered distributions, but this presents no essential difficulty. With the background on generalized functions presented in Chaps. 2 and 3, we should have no qualms about issues of existence of the transform.

On the other hand, the Fourier transform of a random process is another random process, and we are usually more interested in averages than in properties of individual samples. In particular, with finite-power processes, we often want to know how the average power is distributed as a function of frequency. The branch of stochastic theory that addresses this question is called *spectral analysis*, and a frequency-domain description of the average power is known as a *spectrum*, *power spectrum* or *power spectral density*.

We shall give a brief overview of the historical development of spectral analysis and then give two equivalent definitions of power spectral density. Initially the discussion will consider stationary processes in the time domain, but then we make the transition to the space domain as we see how the theory can be applied to processes that are not exactly stationary.

A brief history of spectra The early history of spectral analysis was motivated by a desire to understand white light (Gouy, 1886; Rayleigh, 1903; Schuster, 1894, 1904, 1906). Gouy's work was based on the Fourier series, while Lord Rayleigh used the newly developed Plancherel ($\mathbb{L}_2$) interpretation of the Fourier transform. Wiener (1930) marvels (though not without a touch of irony) at these forays: "In both cases one is astonished by the skill with which the authors use clumsy and unsuitable tools to obtain the right results, and one is led to admire the unfailing heuristic insight of the true physicist."

Wiener's own pioneering treatise, *Generalized Harmonic Analysis* (Wiener, 1930), was built on the work of Sir Arthur Schuster. Schuster used a windowed or truncated function defined by

$$f_T(t) = f(t)\,\mathrm{rect}(t/T)\,, \tag{8.126}$$

with a Fourier transform defined by

$$F_T(\nu) = \int_{-\frac{1}{2}T}^{\frac{1}{2}T} dt\ f(t) \exp(-2\pi i \nu t) \,. \tag{8.127}$$

Schuster proposed specifying the spectrum of $f(t)$ by the *periodogram*, defined by

$$S_p(\nu) = \lim_{T\to\infty} \frac{1}{T} |F_T(\nu)|^2 \,. \tag{8.128}$$

By (3.135), $|F_T(\nu)|^2$ is the Fourier transform of the deterministic autocorrelation *integral* (not to be confused with the statistical autocorrelation function) of $f_T(t)$. Thus (8.128) is equivalent to

$$S_p(\nu) = \lim_{T\to\infty} \frac{1}{T} \boldsymbol{\mathcal{F}} \{ [f_T \star f_T^*] \, (x) \} \ , \tag{8.129}$$

where $\boldsymbol{\mathcal{F}}$ is the Fourier operator and, by (3.115),

$$[f_T \star f_T^*] \, (t) = \int_{-\infty}^{\infty} dt'\, f_T(t+t')\, f_T^*(t') \,. \tag{8.130}$$

Wiener's approach was slightly different. He defined

$$R_{W,T}(t) = \frac{1}{T} \int_{-\frac{1}{2}T}^{\frac{1}{2}T} dt'\, f(t+t')\, f^*(t') \,, \tag{8.131}$$

which differs from (8.130) mainly in the fact that the truncation is on the limits rather than on both functions separately; there is also a factor of $1/T$ built into the definition.

The only requirement placed on the function $f(t)$ is that $R_{W,T}(t) < \infty$ for all t, but this turns out to be a very useful mathematical condition (Champeney, 1987). The special case $t = 0$ shows that these functions must be finite-power functions as defined in (8.70). For such functions, Wiener defined a spectrum by

$$S_W(\nu) = \lim_{T\to\infty} \boldsymbol{\mathcal{F}} \{ R_{W,T}(t) \} \,. \tag{8.132}$$

Note that neither S_p nor S_W involves any statistical average; both Wiener and Schuster took a functional or deterministic viewpoint and did not invoke ensembles of any kind. Thus their spectra apply to a single realization of the random process, albeit one of infinite length. For any function for which S_W is finite, S_W and S_p are identical (Champeney, 1987).

Convergence issues In practice, one might think that a reasonable approximation of S_p or S_W could be obtained by using a single periodogram of finite length and just omitting the limit $T \to \infty$ in (8.128) or (8.132). It might also be expected that this approximation would get better as T gets larger. In fact, however, the Fourier transform of a single sample function of a random process is a very poor spectral measure.

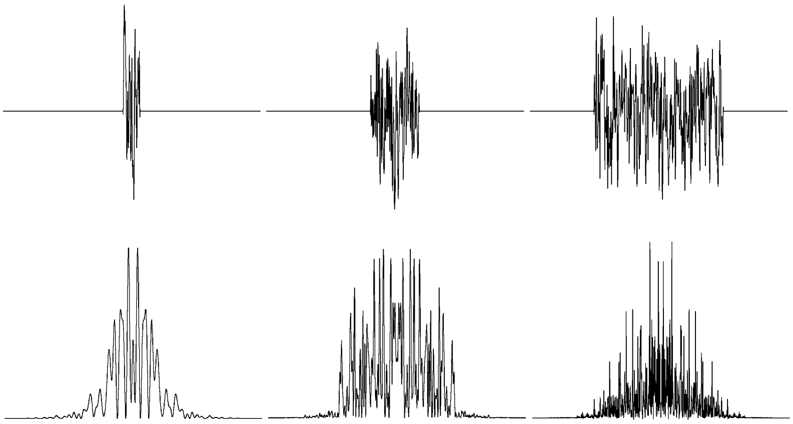


Fig. 8.1 Three sample functions of a random process (top) and their periodograms (bottom). The random process was created by calling a uniform random-number generator independently at each of 1024 sample points, then performing a discrete convolution with a Gaussian to produce a random process with a Gaussian power spectrum. The sample functions were windowed as shown, and the periodograms were computed by discrete Fourier transforms.

This point is illustrated in Fig. 8.1, which shows three sample functions of different length of a stationary random process, along with the corresponding finite-length periodograms. Note that the periodograms do not smoothly approach a limit as $T \to \infty$ but instead oscillate ever more rapidly.

One way to deal with the rapid oscillation is to average the periodogram by convolution with some smooth function. In fact, this average can be built in by windowing the samples with the Fourier transform of the smoothing function. This approach smooths out any fine details that might be present in the spectrum but provides better convergence as T gets large. Some additional approaches to this problem will be discussed briefly in Sec. 8.4.4.

Power spectra as statistical averages Another way to fix the convergence problems associated with S_p and S_W is to use not one but many independent realizations of the random process and to average the resulting periodograms. In the limit of an infinite number of realizations, this approach, pioneered by Khinchin, amounts to incorporating a statistical average in the definition of the spectrum. Khinchin's definition was

$$S_{ac}(\nu) = \mathcal{F}\{R(\Delta t)\} = \int_{-\infty}^{\infty} d\Delta t \ \langle f(t+\Delta t)\, f^*(t)\rangle \exp(-2\pi i\nu\Delta t)\,, \tag{8.133}$$

where the subscript *ac* indicates that this version of the spectrum is derived from the autocorrelation function $R(\Delta t)$ of a stationary random process. The spectrum defined this way is well behaved mathematically and universally used. Equation (8.133) is often referred to as the *Wiener-Khinchin theorem*, though it is really a definition rather than a theorem.

Expected periodogram Another way to incorporate an ensemble average into the definition of the spectrum is to take the expectation of the periodogram,

$$S_{ep}(\nu) = \lim_{T\to\infty} \frac{1}{T} \left\langle |F_T(\nu)|^2 \right\rangle . \tag{8.134}$$

Unlike $S_{ac}(\nu)$, $S_{ep}(\nu)$ is defined for nonstationary as well as stationary random processes, though they have to be finite-power processes for S_{ep} to be nonzero. For stationary processes, however, $S_{ep}(\nu)$ is equivalent to $S_{ac}(\nu)$, as we shall now show.

From the definition of $F_T(\nu)$, we can write

$$S_{ep}(\nu) = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} dt' \; \langle f(t)\, f^*(t')\rangle \,\mathrm{rect}\left(\frac{t}{T}\right) \mathrm{rect}\left(\frac{t'}{T}\right) \exp[2\pi i(t'-t)\nu]\,. \tag{8.135}$$

Now we make the change of variables $(t, t') \to (t, \Delta t)$, where $\Delta t = t - t'$. With the assumption that $\langle f(t)\, f^*(t')\rangle = R(\Delta t)$ and a little algebra, we find

$$S_{ep}(\nu) = \lim_{T\to\infty} \int_{-\infty}^{\infty} d\Delta t \; R(\Delta t)\,\mathrm{tri}\left(\frac{\Delta t}{T}\right) \exp(-2\pi i\nu\Delta t)\,, \tag{8.136}$$

where the function $\mathrm{tri}(\cdot)$ is defined in (3.139).

We can now use the convolution theorem (3.132) along with (3.142) to write

$$S_{ep}(\nu) = \lim_{T\to\infty} S_{ac}(\nu) * T\,\mathrm{sinc}^2(T\nu)\,. \tag{8.137}$$

But we know from (2.87) that $T\,\mathrm{sinc}^2(T\nu)$ is a valid limiting representation of $\delta(\nu)$. From Sec. 3.3.6 we also know that convolution of $S_{ac}(\nu)$ with $\delta(\nu)$ reproduces $S_{ac}(\nu)$ if that function is either a good function (defined in Sec. 2.1.2) or a generalized function of compact support (defined in Sec. 3.3.6). The support can be chosen arbitrarily large, or we can argue as in Sec. 2.3.1 that any generalized function can be approximated arbitrarily closely by a good function.

Thus, with essentially no restrictions beyond stationarity, we have

$$S_{ep}(\nu) = S_{ac}(\nu)\,. \tag{8.138}$$

Because of this equivalence, we shall delete the subscripts henceforth and denote the power spectral density simply by $S(\nu)$. Either definition, (8.133) or (8.134), will be used as convenient.

Spatial power spectra Stationary spatial random processes were discussed in Sec. 8.2.4. If this model is used, the spatial version of the Wiener-Khinchin theorem, (8.133), is

$$S(\boldsymbol{\rho}) = \int_{\infty} d^q\Delta r \; R(\Delta\mathbf{r}) \exp(-2\pi i\boldsymbol{\rho}\cdot\Delta\mathbf{r})\,. \tag{8.139}$$

Stochastic Wigner distribution function A general way of applying Fourier analysis to nonstationary random processes is to make use of the Wigner distribution function, defined in Sec. 5.2.1. For a spatial random process $f(\mathbf{r})$, we define the stochastic Wigner function by [*cf.* (5.54)]

$$W_f(\mathbf{r}_0, \boldsymbol{\rho}) = \int_{\infty} d^q\Delta r \; \left\langle f(\mathbf{r}_0 + \tfrac{1}{2}\Delta\mathbf{r})\, f^*(\mathbf{r}_0 - \tfrac{1}{2}\Delta\mathbf{r})\right\rangle \exp(-2\pi i\boldsymbol{\rho}\cdot\Delta\mathbf{r})\,. \tag{8.140}$$

This expression should be compared to the Wiener-Khinchin theorem for a stationary random process, (8.139), which can be written in symmetrized form as

$$S(\boldsymbol{\rho}) = \int_\infty d^q\Delta r \ \langle f(\mathbf{r}+\Delta\mathbf{r})\, f^*(\mathbf{r})\rangle \exp(-2\pi i\boldsymbol{\rho}\cdot\Delta\mathbf{r})$$

$$= \int_\infty d^q\Delta r \ \left\langle f(\mathbf{r}_0+\tfrac{1}{2}\Delta\mathbf{r})\, f^*(\mathbf{r}_0-\tfrac{1}{2}\Delta\mathbf{r})\right\rangle \exp(-2\pi i\boldsymbol{\rho}\cdot\Delta\mathbf{r})\,, \tag{8.141}$$

where the second equality follows since the autocorrelation function is independent of shifts of the coordinate system for a stationary process. Thus, if the process is stationary, the stochastic Wigner function is independent of $\mathbf{r}_0$ and is precisely the power spectral density.

For nonstationary processes, however, $W_f(\mathbf{r}_0,\boldsymbol{\rho})$ is a function of $\mathbf{r}_0$ as well as $\boldsymbol{\rho}$; it can be interpreted as the spectral content associated with point $\mathbf{r}_0$. This interpretation is reinforced by examining the quasistationary case. From (8.119) and (8.140) we can write

$$W_f(\mathbf{r}_0,\boldsymbol{\rho}) = b(\mathbf{r}_0)\int_\infty d^q\Delta r \ a(\Delta\mathbf{r})\exp(-2\pi i\boldsymbol{\rho}\cdot\Delta\mathbf{r}) = b(\mathbf{r}_0)\,A(\boldsymbol{\rho})\,. \tag{8.142}$$

Here the Wigner distribution function is just the Fourier transform of the short-range part of the autocorrelation function, modulated by the shift-variant strength of the slowly varying component at $\mathbf{r}_0$.

8.2.6 Linear filtering of random processes

We now derive the autocorrelation function of the output process that results from linear filtering of a given random process. We shall consider stationary and nonstationary random processes and shift-invariant and shift-variant filters.

Nonstationary process, shift-variant filter We first consider the case where a random process $g(\mathbf{r})$ is generated as the output of the transformation of an input random process $f(\mathbf{r})$ by a linear shift-variant filter whose impulse response is denoted $h(\mathbf{r},\mathbf{r}')$. The output of the filter at positions $\mathbf{r}$ and $\mathbf{r}+\Delta\mathbf{r}$ can be written, respectively, as

$$g(\mathbf{r}) = \int_\infty d^q r'\, h(\mathbf{r},\mathbf{r}')\, f(\mathbf{r}')\,, \tag{8.143}$$

$$g(\mathbf{r}+\Delta\mathbf{r}) = \int_\infty d^q r'\, h(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}')\, f(\mathbf{r}')\,. \tag{8.144}$$

By direct substitution of these expressions into the definition, (8.97), we obtain for the autocorrelation of the output process at positions $\mathbf{r}$ and $\mathbf{r}+\Delta\mathbf{r}$:

$$R_g(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}) = \langle g(\mathbf{r}+\Delta\mathbf{r})\, g^*(\mathbf{r})\rangle$$

$$= \left\langle \int_\infty d^q r'\, h(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}')\, f(\mathbf{r}') \int_\infty d^q r''\, h^*(\mathbf{r},\mathbf{r}'')\, f^*(\mathbf{r}'')\right\rangle$$

$$= \int_\infty d^q r' \int_\infty d^q r''\, h(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}')\, R_f(\mathbf{r}',\mathbf{r}'')\, h^*(\mathbf{r},\mathbf{r}'')\,. \tag{8.145}$$

The corresponding expression for the autocovariance is

$$K_g(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}) = R_g(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}) - \langle g(\mathbf{r}+\Delta\mathbf{r})\rangle \langle g^*(\mathbf{r})\rangle$$

$$= \int_\infty d^q r' \int_\infty d^q r''\, h(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}')\, K_f(\mathbf{r}',\mathbf{r}'')\, h^*(\mathbf{r},\mathbf{r}'')\,. \qquad (8.146)$$

This is the most general form for the autocovariance after linear filtering. It is the continuous analog of the discrete result given in (8.50), as one can see by rewriting it in operator form:

$$\boldsymbol{\mathcal{K}}_{\mathbf{g}} = \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_{\mathbf{f}}\boldsymbol{\mathcal{H}}^\dagger\,, \qquad (8.147)$$

where $\boldsymbol{\mathcal{K}}_{\mathbf{f}}$ is the autocovariance operator, *i.e.*, the integral operator with kernel $K_{\mathbf{f}}(\mathbf{r},\mathbf{r}')$, and similarly for $\boldsymbol{\mathcal{K}}_{\mathbf{g}}$, while $\boldsymbol{\mathcal{H}}$ describes the filter. There are no restrictions on $\boldsymbol{\mathcal{H}}$ in this equation, except that it must be a linear operator. It even applies to linear CD operators, though in that case the left-hand side is a covariance matrix rather than an autocovariance operator.

Nonstationary process, shift-invariant filter We consider next the case where the random process $g(\mathbf{r})$ is generated as the output of the transformation of a general input random process $f(\mathbf{r})$ by a linear shift-invariant filter with impulse response $h(\mathbf{r})$. The processes $g(\mathbf{r})$ and $f(\mathbf{r})$ are now related by convolution:

$$g(\mathbf{r}) = \int_\infty d^q r'\, h(\mathbf{r}-\mathbf{r}')\, f(\mathbf{r}') = h(\mathbf{r}) * f(\mathbf{r})\,, \qquad (8.148)$$

where the notation of Sec. 3.3.6 has been used.

We can obtain the autocorrelation of the output process $g(\mathbf{r})$ from that of the input process $f(\mathbf{r})$ by substituting (8.148) into (8.145):

$$R_g(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}) = \langle g(\mathbf{r}+\Delta\mathbf{r})\, g^*(\mathbf{r})\rangle$$

$$= \left\langle \int_\infty d^q r'\, h(\mathbf{r}+\Delta\mathbf{r}-\mathbf{r}')\, f(\mathbf{r}') \int_\infty d^q r''\, h^*(\mathbf{r}-\mathbf{r}'')\, f^*(\mathbf{r}'') \right\rangle . \qquad (8.149)$$

Alternatively, we have

$$R_g(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}) = \left\langle \int_\infty d^q r'\, h(\mathbf{r}')\, f(\mathbf{r}+\Delta\mathbf{r}-\mathbf{r}') \int_\infty d^q r''\, h^*(\mathbf{r}'')\, f^*(\mathbf{r}-\mathbf{r}'') \right\rangle$$

$$= \int_\infty d^q r'\, h(\mathbf{r}') \int_\infty d^q r''\, h^*(\mathbf{r}'')\, R_f(\mathbf{r}+\Delta\mathbf{r}-\mathbf{r}',\mathbf{r}-\mathbf{r}'')\,. \qquad (8.150)$$

We can use convolution shorthand to write this equation as

$$R_g(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}) = h(\mathbf{r}+\Delta\mathbf{r}) * R_f(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}) * h^*(\mathbf{r})\,, \qquad (8.151)$$

where the notation indicates that the first convolution is evaluated at the position $\mathbf{r}+\Delta\mathbf{r}$ and the second is evaluated at the position $\mathbf{r}$.

Stationary random process, shift-invariant filter For the special case of a stationary input process, the input correlation function in (8.150) can be written solely in terms of the difference vector as

$$R_f(\mathbf{r}+\Delta\mathbf{r}-\mathbf{r}',\mathbf{r}-\mathbf{r}'') = \langle f(\mathbf{r}+\Delta\mathbf{r}-\mathbf{r}')\, f^*(\mathbf{r}-\mathbf{r}'')\rangle = R_f(\Delta\mathbf{r}-\mathbf{r}'+\mathbf{r}'')\,. \tag{8.152}$$

Then (8.150) can be written

$$R_g(\Delta\mathbf{r}) = \int_\infty d^q r'\, h(\mathbf{r}') \int_\infty d^q r''\, h^*(\mathbf{r}'')\, R_f(\Delta\mathbf{r}-\mathbf{r}'+\mathbf{r}'')\,. \tag{8.153}$$

This equation is often written in a shorthand notation as (Papoulis, 1965)

$$R_g(\Delta\mathbf{r}) = \langle g(\mathbf{r}+\Delta\mathbf{r})\, g^*(\mathbf{r})\rangle = h(\Delta\mathbf{r}) * R_f(\Delta\mathbf{r}) * h^*(-\Delta\mathbf{r})\,. \tag{8.154}$$

This notation refers to the fact that the first operation is an ordinary convolution, but the second is actually a correlation. In this shorthand a correlation is written using the convolution notation with a change of sign of the argument. Alternatively, one can use $\star$ to represent the correlation integral:

$$R_g(\Delta\mathbf{r}) = \langle g(\mathbf{r}+\Delta\mathbf{r})\, g^*(\mathbf{r})\rangle = [h * R_f \star h^*]\,(\Delta\mathbf{r})\,. \tag{8.155}$$

Fourier transformation of (8.155) yields the important formula

$$S_g(\boldsymbol{\rho}) = S_f(\boldsymbol{\rho})\, |H(\boldsymbol{\rho})|^2\,, \tag{8.156}$$

where $H(\boldsymbol{\rho}) = \mathcal{F}_q\{h(\mathbf{r})\}$. Thus, when a stationary random process is filtered by a linear shift-invariant filter, the power spectral density on the output of the filter is the power spectral density on the input times the squared modulus of the filter transfer function. This result should be compared to the familiar result for shift-invariant filtering of a deterministic signal. From (3.132) we know that

$$G(\boldsymbol{\rho}) = H(\boldsymbol{\rho})\, F(\boldsymbol{\rho})\,. \tag{8.157}$$

In the context of stationary random processes, (8.157) applies to *individual sample functions* while (8.156) applies to the power spectral densities.

Filtering of delta-correlated processes We are often concerned with random processes where the correlation has such short range that $R_{\Delta f}(\mathbf{r},\mathbf{r}')$ can be approximated by $b(\mathbf{r})\,\delta(\mathbf{r}-\mathbf{r}')$. A prime example, the Poisson random process, will be discussed in detail in Chap. 11. Another example is *white noise*, a stationary process that has a flat power spectrum and hence a delta-function correlation. We now investigate the effect of linear filtering on delta-correlated processes.

With delta correlation, the general space-variant filter equation, (8.144), leads to

$$\begin{aligned} R_{\Delta g}(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}) = \int_\infty d^q r'\, h(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}') \int_\infty d^q r''\, b(\mathbf{r}')\, \delta(\mathbf{r}'-\mathbf{r}'')\, h^*(\mathbf{r},\mathbf{r}'') \\ \int_\infty d^q r'\, h(\mathbf{r}+\Delta\mathbf{r},\mathbf{r}')\, b(\mathbf{r}')\, h^*(\mathbf{r},\mathbf{r}')\,. \end{aligned} \tag{8.158}$$

For shift-invariant filters, where $h(\mathbf{r}, \mathbf{r}') = h(\mathbf{r} - \mathbf{r}')$, this equation reduces to

$$
\begin{aligned}
R_{\Delta g}(\mathbf{r} + \Delta\mathbf{r}, \mathbf{r}) &= \int_\infty d^q r' \, h(\mathbf{r} + \Delta\mathbf{r} - \mathbf{r}') \, b(\mathbf{r}') \, h^*(\mathbf{r} - \mathbf{r}') \\
&= b(\mathbf{r}) * [h(\mathbf{r} + \Delta\mathbf{r}) \, h^*(\mathbf{r})] \; . \qquad (8.159)
\end{aligned}
$$

The shorthand here requires a brief comment. For purposes of the convolution, the function $[h(\mathbf{r} + \Delta\mathbf{r}) \, h^*(\mathbf{r})]$ is to be regarded as a function of $\mathbf{r}$ for fixed $\Delta\mathbf{r}$. As shown by the integral in (8.158), this product function is then convolved with $b(\mathbf{r})$, and the convolution is repeated for different $\Delta\mathbf{r}$ to get the full dependence of the nonstationary autocorrelation $R_{\Delta g}(\mathbf{r} + \Delta\mathbf{r}, \mathbf{r})$.

Even though $f(\mathbf{r})$ is uncorrelated for any finite lag, the filtering results in a correlation on $g(\mathbf{r})$. Suppose $h(\mathbf{r})$ has a width w in each dimension, *i.e.*, $h(\mathbf{r})$ drops to zero if the magnitude of any component of $\mathbf{r}$ exceeds $\frac{1}{2}w$. Then $[h(\mathbf{r} + \Delta\mathbf{r}) \, h^*(\mathbf{r})]$ drops to zero for all $\mathbf{r}$ if the magnitude of any component of $\Delta\mathbf{r}$ exceeds w. The correlation in $g(\mathbf{r})$ thus has a width in $\Delta\mathbf{r}$ determined by the width of the point spread function.

If $b(\mathbf{r})$ is the constant b_0, so that $R_{\Delta f}(\mathbf{r} - \mathbf{r}') = b_0 \, \delta(\mathbf{r} - \mathbf{r}')$, then we are dealing with stationary white noise and a frequency-domain description is appropriate. The power spectral density of $\Delta f(\mathbf{r})$ is just the constant b_0, and by (8.156) that of $\Delta g(\mathbf{r})$ is given by

$$
S_{\Delta g}(\boldsymbol{\rho}) = b_0 |H(\boldsymbol{\rho})|^2 \, . \qquad (8.160)
$$

The corresponding autocorrelation function is obtained by inverse Fourier transformation:

$$
R_{\Delta g}(\Delta\mathbf{r}) = b_0 [h \star h^*](\Delta\mathbf{r}) \, . \qquad (8.161)
$$

Thus the statistical autocorrelation function for filtered white noise is proportional to the deterministic autocorrelation integral of the impulse response.

8.2.7 Eigenanalysis of the autocorrelation operator

In Sec. 8.1.6, we discussed the eigenvectors and eigenvalues of a covariance matrix. In particular, we showed how a random vector could be expanded in a series with uncorrelated coefficients by using eigenvectors of the covariance matrix as basis vectors. This expansion was called the Karhunen-Loève or KL expansion.

In this section we carry out a similar analysis for a random process, substituting the continuous autocovariance or autocorrelation function for the discrete covariance matrix. One result will be a continuous version of the KL expansion — a linear transformation that will render a correlated process uncorrelated for any finite shift.

To maintain parallelism with Sec. 8.1.6, we restrict attention initially to finite-energy random processes (thus ruling out stationarity), but later we extend the analysis to finite-power processes and in particular to wide-sense stationary ones. In that case we shall find that KL expansion is just Fourier analysis.

Autocorrelation operator It is arbitrary whether we develop KL analysis based on the autocorrelation or autocovariance function; from (8.98) we can easily convert between them. We choose the autocorrelation since we shall eventually make contact with the Wiener-Khinchin theorem (8.133) or (8.139), which defines the power spectral density as the Fourier transform of the autocorrelation function.

For a general, nonstationary, spatial random process $f(\mathbf{r})$, where $\mathbf{r}$ is a qD position vector, the autocorrelation function $R(\mathbf{r},\mathbf{r}')$ is defined by (8.97). For now we restrict attention to square-integrable random processes, so we can regard $R(\mathbf{r},\mathbf{r}')$ as the kernel of an integral operator $\boldsymbol{\mathcal{R}}$ that maps $\mathbb{L}_2(\mathbb{R}^q)$ to itself. Operating on an arbitrary square-integrable function $t(\mathbf{r})$, the operator $\boldsymbol{\mathcal{R}}$ has the form

$$[\boldsymbol{\mathcal{R}}\mathbf{t}\,](\mathbf{r}) = \int_\infty d^q r'\, R(\mathbf{r},\mathbf{r}')\, t(\mathbf{r}')\,. \tag{8.162}$$

Inspection of (8.97) shows that $[R(\mathbf{r},\mathbf{r}')]^* = R(\mathbf{r}',\mathbf{r})$, so $\boldsymbol{\mathcal{R}}$ is Hermitian (see Sec. 1.3.5).

Moreover, as we shall now show, $\boldsymbol{\mathcal{R}}$ is compact. By the discussion in Sec. 1.3.3, an integral operator is compact if its kernel satisfies the Hilbert-Schmidt condition (1.33), which in the present multidimensional case generalizes to

$$\int_\infty d^q r \int_\infty d^q r'\, |\boldsymbol{\mathcal{R}}(\mathbf{r},\mathbf{r}')|^2 < \infty\,. \tag{8.163}$$

Denoting this integral by I_{HS} and inserting (8.97), we can rewrite this condition as

$$I_{HS} = \int_\infty d^q r \int_\infty d^q r'\, |\left\langle f(\mathbf{r})\, f^*(\mathbf{r}')\right\rangle|^2 < \infty\,. \tag{8.164}$$

Now, for any random variable x we know from App. C that $|\langle x\rangle|^2 \le \left\langle |x|^2\right\rangle$. With $x = f(\mathbf{r}')\, f^*(\mathbf{r}')$, this implies that

$$I_{HS} \le \int_\infty d^q r \int_\infty d^q r'\, \left\langle |\,[f(\mathbf{r})\, f^*(\mathbf{r}')]\,|^2\right\rangle\,. \tag{8.165}$$

As discussed in Sec. 8.2.2, we can interchange expectation and integration, yielding

$$I_{HS} \le \left\langle \int_\infty d^q r\, |f(\mathbf{r})|^2 \int_\infty d^q r'\, |f(\mathbf{r}')|^2 \right\rangle\,. \tag{8.166}$$

Every sample function $f(\mathbf{r})$ is assumed to be square-integrable, so each integral in (8.166) is finite. The output of the expectation operation is therefore finite and $I_{HS} \le \infty$. Thus we have shown that $\boldsymbol{\mathcal{R}}$ satisfies the Hilbert-Schmidt condition and is therefore compact.

As discussed in Sec. 1.4.4, a compact Hermitian operator has a denumerable set of eigenfunctions and real eigenvalues. Thus $\boldsymbol{\mathcal{R}}$ satisfies an eigenvalue equation of the form

$$\boldsymbol{\mathcal{R}}\phi_n(\mathbf{r}) = \mu_n \phi_n(\mathbf{r})\,. \tag{8.167}$$

We noted in (8.106) that $\boldsymbol{\mathcal{R}}$ is nonnegative-definite, so $\mu_n \ge 0$. It is convenient to order the eigenvalues by decreasing value:

$$\mu_1 \ge \mu_2 \ge \mu_3 \ge \cdots \ge 0\,. \tag{8.168}$$

Except in very special cases, none of these eigenvalues will be zero, so $\boldsymbol{\mathcal{R}}$ has infinite rank.

Karhunen-Loève expansions Since the eigenfunctions of a Hermitian operator can be chosen to form an orthonormal basis, any function $f(\mathbf{r})$ in the domain of $\boldsymbol{\mathcal{R}}$, *i.e.*, $\mathbb{L}_2(\mathbb{R}^q)$, can be expanded in the form

$$f(\mathbf{r}) = \sum_{n=1}^{\infty} \alpha_n \phi_n(\mathbf{r}) \,, \tag{8.169}$$

where the coefficients are given by scalar products of the form

$$\alpha_n = (\phi_n(\mathbf{r}), f(\mathbf{r})) \,. \tag{8.170}$$

If $f(\mathbf{r})$ is a sample function of a random process, then the coefficients α_n are random variables. If $f(\mathbf{r})$ is drawn from the ensemble described by $\boldsymbol{\mathcal{R}}$, then these coefficients are uncorrelated, as we shall now demonstrate. The cross-correlation of two coefficients, α_n and α_m, is given by

$$\langle \alpha_n \alpha_m^* \rangle = \langle (\phi_n(\mathbf{r}), f(\mathbf{r}))\, (\phi_m(\mathbf{r}'), f(\mathbf{r}'))^* \rangle \;. \tag{8.171}$$

Writing out the scalar products as integrals and again interchanging integration and expectation, we find

$$\langle \alpha_n \alpha_m^* \rangle = \int_\infty d^q r \int_\infty d^q r' \; \phi_n^*(\mathbf{r})\, \phi_m(\mathbf{r}')\, \langle f(\mathbf{r})\, f^*(\mathbf{r}') \rangle \;. \tag{8.172}$$

By (8.97) and (8.167), we have

$$\langle \alpha_n \alpha_m^* \rangle = \mu_m \int_\infty d^q r \;\, \phi_n^*(\mathbf{r})\, \phi_m(\mathbf{r}) \,, \tag{8.173}$$

and the orthonormality of the eigenfunctions yields, finally,

$$\langle \alpha_n \alpha_m^* \rangle = \mu_n \, \delta_{nm} \,. \tag{8.174}$$

Thus the expansion in (8.169) generalizes the Karhunen-Loève expansion of random vectors, as discussed in Sec. 8.1.6, to random processes.

Stationary random processes The derivation above of the KL expansion is not directly applicable to stationary random processes since their sample functions are not square-integrable. Hence the autocorrelation operator is not compact and its eigenvalues are not denumerable.

Since the discrete index n on $\phi_n(\mathbf{r})$ and μ_n is no longer appropriate, we shall leave off any index until we discover what to use. The eigenvalue equation for a stationary random process is then

$$\int_\infty d^q r' \, R(\mathbf{r} - \mathbf{r}')\, \phi(\mathbf{r}') = \mu\, \phi(\mathbf{r}) \,. \tag{8.175}$$

A simple change of variables yields

$$\int_\infty d^q r' \, R(\mathbf{r}')\, \phi(\mathbf{r} - \mathbf{r}') = \mu\, \phi(\mathbf{r}) \,. \tag{8.176}$$

Direct substitution shows that the solution of this equation is

$$\phi(\mathbf{r}) = \exp(2\pi i \boldsymbol{\rho} \cdot \mathbf{r})\,, \tag{8.177}$$

$$\mu = \int_\infty d^q r'\, R(\mathbf{r}') \exp(-2\pi i \boldsymbol{\rho} \cdot \mathbf{r}') = \boldsymbol{\mathcal{F}}_q\{R(\mathbf{r})\} = S(\boldsymbol{\rho})\,, \tag{8.178}$$

where $S(\boldsymbol{\rho})$ is the power spectral density as defined in (8.139). Thus, for a stationary random process, the eigenfunctions of the autocorrelation operator are Fourier basis functions (or plane waves), and the eigenvalues are given by the power spectral density. The problem is mathematically equivalent to singular-value decomposition of a linear, shift-invariant system as discussed in Sec. 7.2.5

The eigenfunctions and eigenvalues are distinguished by a continuous vector index $\boldsymbol{\rho}$ (the spatial frequency), rather than by a discrete index n. Thus we denote the eigenfunction in (8.177) as $\phi_{\boldsymbol{\rho}}(\mathbf{r})$ and the eigenvalue as $\mu_{\boldsymbol{\rho}}$. With this notation, the KL expansion (8.169) becomes

$$f(\mathbf{r}) = \int_\infty d^q\rho\; F(\boldsymbol{\rho})\, \phi_{\boldsymbol{\rho}}(\mathbf{r}) = \int_\infty d^q\rho\; F(\boldsymbol{\rho}) \exp(2\pi i \boldsymbol{\rho} \cdot \mathbf{r})\,. \tag{8.179}$$

By analogy with (8.170), the expansion coefficients $F(\boldsymbol{\rho})$ are given by

$$F(\boldsymbol{\rho}) = \int_\infty d^q r\; f(\mathbf{r}) \exp(-2\pi i \boldsymbol{\rho} \cdot \mathbf{r})\,. \tag{8.180}$$

Formally, (8.179) states that the KL expansion is simply the representation of a sample function of the stationary random process by its inverse Fourier transform, while (8.180) says that the expansion coefficient is the Fourier transform of the sample function. In this sense, KL expansion reduces to Fourier analysis in the stationary case. In a strict mathematical sense, however, this interpretation raises some problems. If $f(\mathbf{r})$ is a sample function from a stationary random process, it must have the same mean value at all points in the infinite domain $\mathbb{R}^q$. Hence it is not square-integrable or absolutely integrable, and the classical Fourier existence and convergence theorems do not apply.

We can fix these problems in one of two ways. One approach is to presume that each sample function is truncated by a window function of finite size, and then let this size go to infinity as in Sec. 8.2.5. A neater approach is simply to regard $f(\mathbf{r})$ as a generalized function related to a tempered distribution. This requires only that the sample function be integrable when multiplied by a good function such as a Gaussian, which is an easy condition to satisfy. From the discussion in Sec. 3.3.4, we know that $F(\boldsymbol{\rho})$ is also a generalized function in that case. For example, if $f(\mathbf{r})$ has a nonzero mean $\overline{f}$ (which must be independent of $\mathbf{r}$ because of the stationarity), then $F(\boldsymbol{\rho})$ must contain a term $\overline{f}\,\delta(\boldsymbol{\rho})$.

From the viewpoint of generalized functions, we can now discuss the correlation properties of the expansion coefficients $F(\boldsymbol{\rho})$. A derivation paralleling the one that led to (8.173) shows that

$$\langle F(\boldsymbol{\rho})\, F^*(\boldsymbol{\rho}')\rangle = S(\boldsymbol{\rho})\, \delta(\boldsymbol{\rho} - \boldsymbol{\rho}')\,. \tag{8.181}$$

Just as in (8.173), the KL expansion coefficients are orthogonal for a stationary random process, but now orthogonality is defined with a Dirac delta rather than a Kronecker delta. Thus Fourier transformation of a stationary random process results in a delta-correlated random process. We shall make use of this result in the next chapter on Poisson random processes.

Another important conclusion from (8.181) is that the second moment $\left\langle |F(\boldsymbol{\rho})|^2 \right\rangle$ is infinite for a stationary random process. Since the mean of the Fourier transform, $\langle F(\boldsymbol{\rho})\rangle$, is the same as the Fourier transform of the mean, $\boldsymbol{\mathcal{F}}\{\langle f(\mathbf{r})\rangle\}$, we would not expect $|\langle F(\boldsymbol{\rho})\rangle|^2$ to be infinite (except possibly for $\boldsymbol{\rho} = 0$), so (8.181) implies that the variance of the Fourier transform of a stationary random process is also infinite.

8.2.8 Discrete random processes

As we discussed in Chap. 7, digital images are discrete vectors, and it is often useful to model actual, physical objects as discrete vectors also. When we analyze the stochastic properties of digital images or discrete object models, then, they become random vectors. The general treatment of random vectors from Sec. 8.1 is applicable here, but there is also an additional structure we can exploit. If a random vector **g** represents an image and each component of the vector represents a pixel, we are interested above all in the relationship between the values at different pixels. If we shuffled the pixels into a different arrangement, they would not represent the same image.

A similar situation occurs in discussing random temporal signals, where the temporal ordering of the signal values is key. For example, if a random analog waveform $f(t)$ is sampled at regular time points for further digital processing, the sequence of values $\{f(t_n)\}$ constitutes a random vector in which the order of the elements must be maintained.

We shall use the term *discrete random process*[5] to mean a random vector in which crucial information is contained in the temporal or spatial arrangement of the component values. Loosely, a discrete random process is a random vector endowed with a topology. For temporal processes, the term *random sequence* is often used, and some books adopt this term for the spatial case as well.

Discrete stationarity in 1D Suppose the sequence $\{f_n\}$ is obtained by sampling a stationary temporal random process $f(t)$ at regular intervals $t = t_n = n\Delta t$. The sampling could be simple point sampling where $f_n = f(t_n)$, but a more general form is

$$f_n = \int_{-\infty}^{\infty} dt\; f(t)\, s(t_n - t)\,. \tag{8.182}$$

The sampling function $s(t)$ is a delta function for point sampling, but in general it is unrestricted in what follows. Note that (8.182) is in the form of a convolution, so f_n consists of point samples of the random process $[f * s](t)$.

If $f(t)$ is wide-sense stationary, so is $[f * s](t)$. It then follows that the covariance matrix of the samples $\{f_n\}$ satisfies [*cf.* (8.112)]

$$K_{nn'} = k_{n-n'}\,. \tag{8.183}$$

Note that the left-hand side of this equation has two indices but the right-hand side has just one; if there are N elements in the sequence $\{f_n\}$, there are N^2 elements

---

[5] Note that the elements of the random vector need not be discrete random variables; the term *discrete* here refers to the temporal or spatial variable.

in the matrix $\mathbf{K}$ but only N independent ones. Each row of the matrix is a shifted version of every other row. Matrices with this structure are said to be *Toeplitz.*

Circulant covariance matrices We encountered Toeplitz matrices in a deterministic context in Chap. 7. Specifically, we saw in Sec. 7.4.4 that a considerable mathematical simplification resulted if we could approximate the Toeplitz matrix by a circulant one, where the difference $n - n'$ in (8.183) is interpreted modulo N, with N being the total number of samples. For example, if n and n' run from 0 to 255, then the pairs $(n = 10, n' = 5)$ and $(n = 2, n' = 253)$ have the same value for $n - n'$ modulo 256 and hence the same correlation if $\mathbf{K}$ is a 256×256 circulant matrix. Physically, of course, this makes no sense; elements 5 and 10 of the sequence are close together and might be expected to be correlated, but elements 2 and 253 are widely separated, and there is no reason to believe that they should have the same correlation as elements 5 and 10.

Nevertheless, the circulant approximation to a Toeplitz covariance matrix is often used, just as is the circulant approximation to a discrete convolution [see Sec. 7.4.4, especially (7.344)]. The error might be tolerable if the kernel ($k_{n-n'}$ in the stochastic problem or h_{m-n} in the deterministic problem) is compact and our interest does not extend to the extreme elements in the sequence. Some vigilance is required to be sure that we do not fall into a trap when we assume that a Toeplitz matrix is approximately circulant.

The reason we might want to make this approximation was laid out in Sec. 7.4.4: a circulant matrix is diagonalized by a DFT [see (7.352)]. For the deterministic DD problem considered in Sec. 7.4.4, that meant that the DFT basis was essentially the SVD basis when the system was described by a circulant $\mathbf{H}$ matrix. In the stochastic context of this chapter, the DFT basis is the KL basis when we can use the circulant form for the covariance.

Discrete spatial stationarity Circulant stationarity is even more suspect than continuous stationarity in imaging applications, but for completeness we state the mathematical results explicitly. If we consider an image $\mathbf{g}$ to be a qD discrete random process, then the elements of the image can be denoted by $g_{\mathbf{m}}$, where $\mathbf{m}$ is a qD multi-index as introduced in Sec. 3.4.6. If each component m_i of $\mathbf{m}$ runs from 0 to $M-1$, then circulant stationarity means that $[\mathbf{K_g}]_{\mathbf{mm}'}$ depends on $m_i - m_i'$ modulo M for all i. In that case, as discussed in Sec. 7.4.4, the circulant covariance matrix is diagonalized by a qD DFT, and the basis vectors in this transform comprise the KL basis.

The cyclic character of the covariance matrix becomes less objectionable as the array gets larger if the correlation length is constant. In the limit as $M \to \infty$, the distinction between Toeplitz and circulant vanishes. In that case, the Toeplitz/circulant matrix is diagonalized by the discrete-space Fourier transform (DSFT) introduced in Sec. 3.6.4, and the KL basis vectors form a continuous basis indexed by the spatial-frequency vector $\boldsymbol{\rho}$. To use this basis, however, we must now make two unphysical assumptions: an infinite amount of data and discrete stationarity over an infinite domain.

8.3 NORMAL RANDOM VECTORS AND PROCESSES

Among the many probability laws for continuous random variables, the normal probability law is certainly the most commonly encountered. The fundamental reason for this is that when statistically independent random variables are added together, their sum asymptotically follows the normal distribution. (We shall provide a more rigorous treatment of this principle later in this section.) The second reason for the popularity of the normal law is that, as we shall soon see, its structure leads to straightforward and well-understood manipulations. The third reason follows from the first two: a great collection of practically useful statistical tools develop as elaborations upon the normal probability law.

The normal law is frequently named for C. F. Gauss (1777–1855), whose *Theory of the Combination of Observations* (1823) has earned him this eponymity. We shall use the terms *normal* and *Gaussian* interchangeably.

8.3.1 Probability density functions

For simplicity we consider here only real random variables and vectors, but the complex case is treated in Sec. 8.3.6. The PDF of a real normal random variable g is given (see App. C) by

$$\mathrm{pr}(g) = \left[\frac{1}{2\pi\sigma^2}\right]^{\frac{1}{2}} \exp\left[-\frac{(g-\overline{g})^2}{2\sigma^2}\right] , \tag{8.184}$$

where $\overline{g}$ is the mean of the random variable and σ^2 is its variance. To indicate that a random variable g is drawn from a normal distribution with parameters $\overline{g}$ and σ^2, we write $g \sim \mathcal{N}(\overline{g}, \sigma^2)$.

A multivariate normal random vector is a straightforward generalization of the univariate or scalar case. If each component of an MD random vector $\mathbf{g}$ is a normal random variable, the full probability law on $\mathbf{g}$ is a multivariate normal PDF $\mathrm{pr}(\mathbf{g})$, given by

$$\mathrm{pr}(\mathbf{g}) = \left[(2\pi)^M \det(\mathbf{K})\right]^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{g}-\overline{\mathbf{g}})^t\, \mathbf{K}^{-1}(\mathbf{g}-\overline{\mathbf{g}})\right] , \tag{8.185}$$

where $\overline{\mathbf{g}}$ is the mean vector and $\mathbf{K}$ is the covariance matrix of $\mathbf{g}$ as defined in Sec. 8.1.3. As shown in that section, $\mathbf{K}$ is an $M \times M$, positive-semidefinite Hermitian matrix. The diagonal element K_{mm} of the covariance matrix is the variance of the m^{th} component of $\mathbf{g}$, and the off-diagonal elements of $\mathbf{K}$ are related by $K_{nm} = K_{mn}$ for real vectors. We denote an $M \times 1$ random vector drawn from a multivariate normal distribution with parameters $\overline{\mathbf{g}}$ and $\mathbf{K}$ by $\mathbf{g} \sim \mathcal{N}_M(\overline{\mathbf{g}}, \mathbf{K})$. Its density function is seen from (8.185) to be the exponential of a quadratic form in the random vector.

Diagonalization of the covariance matrix of a Gaussian random vector In Sec. 8.1.6 we showed how the KL expansion of a random vector in terms of the eigenvectors of its covariance matrix results in uncorrelated components. We now revisit the KL expansion procedure for the particular case of Gaussian random vectors. We shall show that, for a multivariate normal, the KL transformation yields a vector with uncorrelated components that are also statistically independent.

From (8.64) we know we can express the inverse of the covariance matrix $\mathbf{K}$ as

$$\mathbf{K}^{-1} = \mathbf{\Phi}\mathbf{M}^{-1}\mathbf{\Phi}^{\dagger}\,, \tag{8.186}$$

where again $\mathbf{\Phi}$ is the matrix formed from the eigenvectors $\boldsymbol{\phi}_m$ of $\mathbf{K}$, and $\mathbf{M}$ is a diagonal matrix with the m^{th} diagonal element equal to the eigenvalue μ_m. We can use (8.186) to rewrite the quadratic form of (8.185) as

$$\begin{aligned}(\mathbf{g}-\overline{\mathbf{g}})^t\mathbf{K}^{-1}(\mathbf{g}-\overline{\mathbf{g}}) &= (\mathbf{g}-\overline{\mathbf{g}})^t\mathbf{\Phi}\mathbf{M}^{-1}\mathbf{\Phi}^{\dagger}(\mathbf{g}-\overline{\mathbf{g}}) \\ &= \left[\mathbf{\Phi}^{\dagger}(\mathbf{g}-\overline{\mathbf{g}})\right]^{\dagger}\mathbf{M}^{-1}\left[\mathbf{\Phi}^{\dagger}(\mathbf{g}-\overline{\mathbf{g}})\right]\,, \end{aligned} \tag{8.187}$$

where we have used the unitarity of $\mathbf{\Phi}$. We define the random vector $\Delta\boldsymbol{\beta}$ by [*cf.* (8.60)]

$$\Delta\boldsymbol{\beta} = \mathbf{\Phi}^{\dagger}(\mathbf{g}-\overline{\mathbf{g}})\,. \tag{8.188}$$

Combining (8.187) and (8.188), we obtain

$$\left[\mathbf{\Phi}^{\dagger}(\mathbf{g}-\overline{\mathbf{g}})\right]^{\dagger}\mathbf{M}^{-1}\left[\mathbf{\Phi}^{\dagger}(\mathbf{g}-\overline{\mathbf{g}})\right] = \Delta\boldsymbol{\beta}^{\dagger}\mathbf{M}^{-1}\,\Delta\boldsymbol{\beta} = \sum_{m=1}^{M}\Delta\beta_m^2/\mu_m\,. \tag{8.189}$$

From (A.73) in App. A, we know that the determinant of $\mathbf{K}$ is the product of its eigenvalues. Using this fact and (8.189), we can rewrite (8.185) as

$$\begin{aligned}\mathrm{pr}(\mathbf{g}) &= (2\pi)^{-M/2}\left[\prod_{m=1}^{M}\mu_m\right]^{-1/2}\exp\left(-\frac{1}{2}\sum_{m=1}^{M}\frac{\Delta\beta_m^2}{\mu_m}\right) \\ &= \prod_{m=1}^{M}(2\pi\mu_m)^{-1/2}\exp\left(-\frac{1}{2}\frac{\Delta\beta_m^2}{\mu_m}\right) = \mathrm{pr}(\boldsymbol{\beta})\,, \end{aligned} \tag{8.190}$$

where the last step is valid since the transformation from $\mathbf{g}$ to $\boldsymbol{\beta}$ is unitary and hence the Jacobian is unity.

Thus, when the quadratic form is diagonalized, the Gaussian multivariate PDF can be written as a product of univariate PDFs, which means that the new variables, $\Delta\beta_m$, are statistically independent. While the components of the random vector $\mathbf{g}$ may covary (as represented by the elements of the covariance matrix $\mathbf{K}$), the components of the random vector $\Delta\beta$ are uncorrelated, with diagonal covariance matrix $\mathbf{M}$, and statistically independent. The mean of each component $\Delta\beta_m$ is 0 and its variance is simply μ_m. The product form of the PDF in (8.190) also makes the normalization of the multivariate Gaussian density readily verifiable.

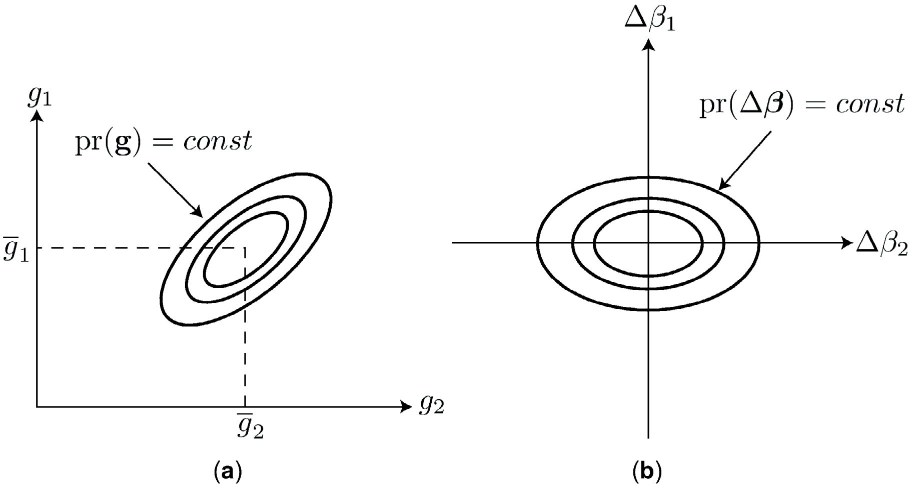

g_1
$\mathrm{pr}(\mathbf{g}) = const$
$\overline{g}_1$
$\overline{g}_2$
g_2
(a)
$\Delta\beta_1$
$\mathrm{pr}(\Delta\boldsymbol{\beta}) = const$
$\Delta\beta_2$
(b)

Fig. 8.2 Contours of constant probability density for a multivariate normal, before and after diagonalization.

Figure 8.2 depicts contours of constant probability for the multivariate normal PDF before and after the diagonalization of $\mathbf{K}$. Following the diagonalization operation the surfaces are found to be ellipsoids whose axes have lengths proportional to the square root of the corresponding eigenvalues μ_m. The diagonalization operation rotates the coordinate axes to coincide with the eigenvectors of $\mathbf{K}$.

When does uncorrelated imply independent? We have just seen that a normal random vector with uncorrelated components also has statistically independent components. The converse always holds — statistically independent components must be uncorrelated — but it is *only* the normal law for which uncorrelated components are statistically independent.

8.3.2 Characteristic function

The diagonalized form of the PDF given in Sec. 8.3.1 provides an easy way to derive the characteristic function of a multivariate normal random vector. From (8.188) and the unitarity of $\mathbf{\Phi}$, we can write $\mathbf{g}$ as

$$\mathbf{g} = \mathbf{\Phi}\Delta\boldsymbol{\beta} + \overline{\mathbf{g}}\,. \tag{8.191}$$

Thus the characteristic function for $\mathbf{g}$ is given by

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \left\langle \exp\left[-2\pi i \boldsymbol{\xi}^t(\mathbf{\Phi}\Delta\boldsymbol{\beta} + \overline{\mathbf{g}})\right]\right\rangle = \exp(-2\pi i \boldsymbol{\xi}^t \overline{\mathbf{g}}) \left\langle \exp\left[-2\pi i (\mathbf{\Phi}^\dagger \boldsymbol{\xi})^t \Delta\boldsymbol{\beta}\right]\right\rangle , \tag{8.192}$$

where we removed a constant factor from the expectation and used the definition of adjoint, (1.39), to get the last form. Using (8.190) for the PDF and writing out the expectation in detail, we find

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \prod_{m=1}^{M} (2\pi\mu_m)^{-1/2} \exp(-2\pi i \xi_m \overline{g}_m)$$
$$\times \int_{-\infty}^{\infty} d\Delta\beta_m \exp\left(-\frac{1}{2}\frac{\Delta\beta_m^2}{\mu_m}\right) \exp\left[-2\pi i (\boldsymbol{\Phi}^\dagger \boldsymbol{\xi})_m \Delta\beta_m\right] . \tag{8.193}$$

Now we have a product of 1D integrals, each of which is just the Fourier transform of a Gaussian; by (3.180) we have

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \prod_{m=1}^{M} \exp(-2\pi i \xi_m \overline{g}_m) \exp\left[-2\pi^2 \mu_m (\boldsymbol{\Phi}^\dagger \boldsymbol{\xi})_m^2\right] . \tag{8.194}$$

From (8.186) we can see that

$$\sum_{m=1}^{M} \mu_m (\boldsymbol{\Phi}^\dagger \boldsymbol{\xi})_m^2 = \boldsymbol{\xi}^t \mathbf{K} \boldsymbol{\xi} , \tag{8.195}$$

so we have, finally,

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \exp(-2\pi i \boldsymbol{\xi}^t \overline{\mathbf{g}}) \exp\left(-2\pi^2 \boldsymbol{\xi}^t \mathbf{K} \boldsymbol{\xi}\right) . \tag{8.196}$$

For $\overline{\mathbf{g}} = \mathbf{0}$, we obtain $\exp(-2\pi^2 \boldsymbol{\xi}^t \mathbf{K} \boldsymbol{\xi})$, which is easy to remember since it is Gaussian in the Fourier domain with spread inverse to that in the domain of the random variable (*i.e.*, $\mathbf{K}$ occurs in place of $\mathbf{K}^{-1}$). The complete form, (8.196), may then be recalled by invoking the Fourier shift theorem (3.108).

Moments We can use the characteristic function given in (8.196) to determine the moments of a multivariate normal random vector. If we apply (8.30) and (8.31) to (8.196), we obtain $\langle \mathbf{g} \rangle = \overline{\mathbf{g}}$ and $\langle \mathbf{g}\, \mathbf{g}^t \rangle = \mathbf{K} + \overline{\mathbf{g}}\overline{\mathbf{g}}^t$. If $\overline{\mathbf{g}} = 0$, then $\langle \mathbf{g}\, \mathbf{g}^t \rangle = \mathbf{K}$. By continuing along this path we find that all odd moments of this distribution are zero for $\overline{\mathbf{g}} = 0$, and all even moments are expressible in terms of $\mathbf{K}$.

We shall find that we frequently need fourth moments of the form $\langle g_i g_j g_k g_l \rangle$, where the g_i, etc., are components of a four-dimensional vector $\mathbf{g}$ distributed as $\mathcal{N}_4(\mathbf{0}, \mathbf{K})$. We can obtain the desired result, referred to as the Gaussian moment theorem, by using the rules for differentiation with respect to a real vector given in Sec. A.9.2. We find that

$$\langle g_i g_j g_k g_l \rangle = \left(\frac{\partial^4 \psi_{\mathbf{g}}(\boldsymbol{\xi})}{\partial \xi_l \partial \xi_k \partial \xi_j \partial \xi_i} \right)_{\boldsymbol{\xi}=\mathbf{0}} = K_{ij} K_{kl} + K_{jk} K_{il} + K_{ik} K_{jl} . \tag{8.197}$$

For the case where $i = j = k = l$, we find $\left\langle g_i^4 \right\rangle = 3\sigma_i^4$, which is a familiar result for univariate normals given in (C.112).

8.3.3 Marginal densities and linear transformations

In this section we derive various descriptors of the behavior of subsets and transformations of the components of a multivariate Gaussian random vector. We start by analyzing the behavior of a single component, regardless of the behavior of the other components, as described by the marginal PDF. We then discuss the behavior of a random vector obtained from linear transformation of a Gaussian random vector.

According to (8.5), the marginal PDF on component g_i of an MD vector $\mathbf{g}$ is obtained by integrating the multivariate PDF over all g_m except for $m = i$. From the central-ordinate theorem of Fourier analysis, (3.104), we know that integrating a function over $(-\infty, \infty)$ is equivalent to setting the frequency to zero in its Fourier transform. Thus the univariate characteristic function for g_i is related to the multivariate characteristic function for $\mathbf{g}$ by

$$\psi_{g_i}(\xi_i) = \psi_{\mathbf{g}}(0, 0, ..., \xi_i, ..., 0)\,. \tag{8.198}$$

With (8.196), we have

$$\psi_{g_i}(\xi_i) = \exp(-2\pi i \xi_i \overline{g}_i) \exp(-2\pi^2 K_{ii} \xi_i^2)\,. \tag{8.199}$$

This is just the characteristic function for a univariate normal with mean $\overline{g}_i$ and variance K_{ii}. Perhaps surprisingly, the form of the marginal on g_i does not depend on K_{im} for $i \neq m$, even though g_i may be correlated with the other components.

Similarly, the bivariate characteristic function for g_i and g_j is given by

$$\begin{aligned} \psi_{g_i,g_j}(\xi_i, \xi_j) &= \psi_{\mathbf{g}}(0, 0, ..., \xi_i, ..., \xi_j, ..., 0) \\ &= \exp\left(-2\pi i \tilde{\boldsymbol{\xi}}^t \tilde{\overline{\mathbf{g}}}\right) \exp\left[-2\pi^2 \tilde{\boldsymbol{\xi}}^t \tilde{\mathbf{K}} \tilde{\boldsymbol{\xi}}\right]\,, \end{aligned} \tag{8.200}$$

where $\tilde{\boldsymbol{\xi}}^t = (\xi_i, \xi_j)$, $\tilde{\overline{\mathbf{g}}} = (\overline{g}_i, \overline{g}_j)^t$ and

$$\tilde{\mathbf{K}} = \begin{bmatrix} K_{ii} & K_{ij} \\ K_{ij} & K_{jj} \end{bmatrix}\,. \tag{8.201}$$

Inverse Fourier transformation of (8.200) yields a bivariate normal PDF with the expected mean and covariance. Again, we do not need to know covariance components other than the ones represented in the marginal of interest.

Other linear transformations of normal random vectors Computation of a marginal is equivalent to finding the PDF for the output of a linear transformation of a random vector. For example, the component g_i can be singled out by computing the scalar product of $\mathbf{g}$ with an $1 \times M$ row vector having a one in the i^{th} column and a zero in all others. Similarly, the 2D vector (g_i, g_j) results from applying a $2 \times M$ matrix operator with ones in positions $(1, i)$ and $(2, j)$ and zeros in all other locations. We now compute the PDF for a random vector formed from a general linear transformation.

Consider the random vector $\mathbf{y} = \mathbf{O}\mathbf{g}$, where $\mathbf{y}$ is a $K \times 1$ vector, $\mathbf{O}$ is a real $K \times M$ matrix and $\mathbf{g} \sim \mathcal{N}_M(\overline{\mathbf{g}}, \mathbf{K})$. The characteristic function for $\mathbf{y}$ follows from (8.43) and (8.196):

$$\psi_{\mathbf{y}}(\boldsymbol{\xi}) = \psi_{\mathbf{g}}(\mathbf{O}^t\boldsymbol{\xi}) = \exp(-2\pi i \boldsymbol{\xi}^t \mathbf{O}\overline{\mathbf{g}}) \exp\left(-2\pi^2 \boldsymbol{\xi}^t \mathbf{O}\mathbf{K}\mathbf{O}^t \boldsymbol{\xi}\right)\,. \tag{8.202}$$

By inspection, then, $\mathbf{y} \sim \mathcal{N}_K(\mathbf{O}\overline{\mathbf{g}}, \mathbf{O}\mathbf{K}\mathbf{O}^t)$. Thus *any* linear transformation of a normal random vector leaves it normal.

In fact, the converse of (8.202) also holds: An $M \times 1$ random vector is normal if and only if its scalar products with all $M \times 1$ vectors are univariate normal (Mardia *et al.*, 1979).

8.3.4 Central-limit theorem

In this section we show that the sum of a large number of random variables tends to be normally distributed. This property, known as the central-limit theorem, is one of the reasons for the prominence of the Gaussian law in probability theory.

We shall introduce the central-limit theorem in stages. Initially we consider *i.i.d.* (independent and identically distributed) scalar random variables, where all moments are finite. These assumptions allow an elementary derivation, though one with restricted validity. Next we discuss the case of i.i.d. random variables where some of the higher moments may be infinite. Then we allow the variables to have different variances and some degree of statistical dependence. Finally we comment briefly on the vector case.

Independent and identically distributed random variables Consider a set of J i.i.d. random variables u_j, $1 \leq j \leq J$, with means $\overline{u}$ and variances σ^2. First we define standardized (zero-mean, unit-variance) random variables by

$$x_j = \frac{u_j - \overline{u}}{\sigma}\,. \tag{8.203}$$

Then we construct a new random variable z, defined by

$$z = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} x_j\,. \tag{8.204}$$

Because the variance of a sum of J i.i.d. random variables is J times the individual variances, and the variance of $x_j/\sqrt{J}$ is $1/J$, z has unit variance. Moreover, since z is a sum of zero-mean random variables, it also has zero mean. We want to show that as $J \to \infty$ the PDF on z tends toward a standard normal distribution, from which it follows readily that the sum of the u_j is normal with mean $J\overline{u}$ and variance $J\sigma^2$.

The derivation proceeds most easily with the aid of characteristic functions. We shall denote the characteristic function of x_j as $\psi_x(\xi)$; no index j is needed since the characteristic function has the same form for all of the x_j. If we assume initially that all moments of x_j are finite, we can expand $\psi_x(\xi)$, in a Taylor series:

$$\begin{aligned}\psi_x(\xi) = \langle \exp(-2\pi i \xi x_j)\rangle &= 1 - 2\pi i \xi \langle x_j \rangle - \frac{4\pi^2}{2!}\xi^2 \left\langle x_j^2 \right\rangle + ... \\ &= 1 - \frac{4\pi^2}{2!}\xi^2 + ...\,,\end{aligned} \tag{8.205}$$

where the second line follows since $\langle x_j \rangle = 0$ and $\left\langle x_j^2 \right\rangle = 1$.

The characteristic function of z is given by

$$\psi_z(\xi) = \langle \exp(-2\pi i \xi z)\rangle = \left\langle \exp\left[-2\pi i \left(\frac{\xi}{\sqrt{J}}\right) \sum_{j=1}^{J} x_j \right]\right\rangle$$

$$= \prod_{j=1}^{J} \left\langle \exp\left[-2\pi i \left(\frac{\xi}{\sqrt{J}}\right) x_j\right]\right\rangle = \prod_{j=1}^{J} \psi_x\left(\frac{\xi}{\sqrt{J}}\right) = \left[\psi_x\left(\frac{\xi}{\sqrt{J}}\right)\right]^J\,, \tag{8.206}$$

where the independence of the x_j has been invoked on the second line to write the expectation of a product as the product of the expectations, and the fact that the x_j are identically distributed is the key to the last step.

We can now insert the Taylor expansion (8.205) into (8.206), yielding

$$\psi_z(\xi) = \left[1 - \frac{2\pi^2\xi^2}{J} + R_J(\xi)\right]^J , \tag{8.207}$$

where $R_J(\xi)$ is the remainder if the Taylor expansion is truncated with the quadratic term. By Taylor's theorem (Rade and Westgren, 1990), $R_J(\xi)$ tends to zero (for any fixed ξ) at least as fast as $J^{-3/2}$ when $J \to \infty$. Thus, in spite of the J^{th} power, these higher terms vanish in the limit. The quadratic term must be retained, however, so that

$$\lim_{J\to\infty} \psi_z(\xi) = \lim_{J\to\infty} \left(1 - \frac{2\pi^2\xi^2}{J}\right)^J = \exp(-2\pi^2\xi^2) , \tag{8.208}$$

which is the characteristic function of a standard-normal random variable. It then follows from the celebrated *continuity theorem* of Paul Lévy (see Loève, 1963) that $z \sim \mathcal{N}(0,1)$.[6]

It is straightforward to go from (8.208) to the probability law for the sum of the original random variables u_j. Defining

$$s_J = \sum_{j=1}^{J} u_j , \tag{8.209}$$

the reader may show that $s_J \sim \mathcal{N}(J\overline{u}, J\sigma^2)$

We have therefore seen that an infinite sum of independent, identically distributed random variables follows a normal distribution, at least when the individual characteristic functions admit of a Taylor expansion. It must be emphasized, however, that the central-limit theorem guarantees normality only asymptotically; it might not be a good approximation for large but finite J. Often the convergence to normality is rapid, requiring as few as perhaps $5-10$ terms, but we should be cautious about finite sums of skewed or otherwise long-tailed PDFs. An extreme example is the case of sums of log-normal distributions, which converge very slowly to the central limit (Barakat, 1976).

Infinite moments There are common PDFs where some of the higher moments are infinite. In Sec. C.5.10, we encountered the Lévy family of distributions, and we noted that the mean was zero but the variance was infinite. A special case of the Lévy distribution is the Cauchy distribution, where $\mathrm{pr}(x) \propto (a^2 + x^2)^{-1}$, a well-known and broadly useful PDF of infinite variance. On the other hand, if we consider $\mathrm{pr}(x) \propto (a^2 + x^2)^{-2}$, then the variance is finite but the fourth moment is infinite. The common feature of these examples is that the characteristic function is not differentiable to all orders and hence cannot be expanded in a Taylor series. Therefore we need to inquire whether it is possible to derive a central-limit theorem.

---

[6] Thanks to Jack Denny for calling our attention to this theorem.

The key is a theorem proved in Shiryayev (1984). If $\langle |x|^n \rangle$ exists for some $n \geq 1$, then the k^{th} derivative of $\psi_x(\xi)$, denoted $\psi_x^{(k)}(\xi)$, exists for every $k \leq n$, and

$$\psi_x(\xi) = \sum_{k=0}^{n} \frac{(2\pi i \xi)^k}{k!} \left\langle x^k \right\rangle + \frac{(2\pi i \xi)^n}{n!} \epsilon_n(\xi) , \tag{8.210}$$

where $|\epsilon_n(\xi)| \leq 3 \langle |x|^n \rangle$ and $\epsilon_n(\xi) \to 0$ as $\xi \to 0$. So long as $\left\langle |x_j|^3 \right\rangle$ is finite, this theorem justifies the steps from (8.205) to (8.208), even when the full Taylor expansion for $\psi_x(\xi)$ does not exist.

For the examples given above, $\left\langle |x_j|^3 \right\rangle$ is infinite for the Lévy and Cauchy PDFs, so the limiting PDF is not normal; in fact, a sum of any number of Lévy random variables is still a Lévy random variable. For $\mathrm{pr}(x) \propto (a^2 + x^2)^{-2}$, however, $\left\langle |x_j|^3 \right\rangle$ is finite and there is a normal central limit.[7]

Independent but not identically distributed random variables Now suppose that the random variables u_j are independent but have different means and variances. Let the mean of u_j be denoted by $\overline{u}_j$ and the variance by σ_j^2, and define

$$x_{jJ} = \frac{u_j - \overline{u}_j}{\sqrt{\sum_{j=1}^{J} \sigma_j^2}} . \tag{8.211}$$

The extra subscript is needed since the denominator depends on J. Note that

$$\langle x_{jJ} \rangle = 0 \qquad \text{and} \qquad \sum_{j=1}^{J} \mathrm{Var}(x_{jJ}) = 1 . \tag{8.212}$$

Now we can define a standardized random variable z by

$$z = \sum_{j=0}^{J} x_{jJ} . \tag{8.213}$$

If the means and variances are independent of j, this definition of z reduces to (8.204).

Shiryayev (1984) discusses various sufficient conditions under which z will tend to a standard normal as $J \to \infty$. They all amount to saying that the variables x_{jJ} are *asymptotically infinitesimal*, in the sense that $\left\langle x_{jJ}^2 \right\rangle \to 0$ as $J \to \infty$, or equivalently that, for every ϵ,

$$\Pr(|x_{jJ}| > \epsilon) \to 0 \qquad \text{as} \qquad J \to \infty . \tag{8.214}$$

This condition is plausible in most practical circumstances because of the denominator in (8.211); so long as the variances σ_j^2 do not themselves tend to zero rapidly as j gets large, the sum of the variances will increase as the number of terms increases, so x_{jJ}, which is normalized by this sum, must get smaller in virtually any sense.

Thus the central-limit theorem states that a sum of asymptotically infinitesimal, zero-mean random variables tends to a standard normal, so long as the sum

---

[7] We thank Dana Clarke for helpful discussions on these examples.

of their variances is normalized to unity (Shiryayev, 1984). From this statement, it is again straightforward to show that the sum of the original variables u_j is also asymptotically normal. Specifically, as $J \to \infty$, s_J becomes distributed as $\mathcal{N}_J[\sum_j \overline{u}_j, \sum_j \mathrm{Var}(u_j)]$.

Sums of dependent random variables Though the central-limit theorem is usually stated for sums of independent random variables, strict independence is not required. For a detailed discussion, see Shiryayev (1984).

Sums of i.i.d. random vectors Central-limit theorems can also be stated for random vectors. We mention here only the simplest case of i.i.d. random vectors where all moments exist.

Let $\mathbf{u}_j$ be an $M \times 1$ random vector with mean $\overline{\mathbf{u}}$ and covariance $\mathbf{K}_{\mathbf{u}}$, both independent of j, and assume that $\mathbf{u}_j$ is independent of $\mathbf{u}_k$ for $j \neq k$. Also let

$$\mathbf{s}_J = \sum_{j=1}^{J} \mathbf{u}_j \,. \tag{8.215}$$

Then, as $J \to \infty$, $\mathbf{s}_J \sim \mathcal{N}_M(J\overline{\mathbf{u}}, J\mathbf{K}_{\mathbf{u}})$. The proof of this statement involves multivariate characteristic functions and the multivariate Taylor expansion (A.179). With this hint, the reader should be able to retrace the steps leading up to (8.208).

8.3.5 Normal random processes

As we shall see in more detail in Sec. 8.4.3, we can sometimes apply the central-limit theorem and argue that the random process representing an object or image is normal. In preparation for that discussion, we examine here some of the mathematical properties of normal random processes. We initially adopt a rather unconventional starting point and define normal random processes in terms of characteristic functionals, but then we shall show that this definition is equivalent to a more common one.

Characteristic functional and linear operators The general form of the characteristic function of a normal random vector is given in (8.196); it can be extended to random processes by use of the characteristic functional, as introduced in Sec. 8.2.3. By analogy to (8.196), we define a real-valued normal random process by requiring that its characteristic functional be given by

$$\Psi_{\mathbf{f}}(\mathbf{s}) = \exp(-2\pi i \mathbf{s}^\dagger \overline{\mathbf{f}}) \exp(-2\pi^2 \mathbf{s}^\dagger \boldsymbol{\mathcal{K}}_{\mathbf{f}} \mathbf{s}) \,, \tag{8.216}$$

where $\boldsymbol{\mathcal{K}}_{\mathbf{f}}$ is the autocovariance operator, *i.e.*, the integral operator with kernel $K_{\mathbf{f}}(\mathbf{r}, \mathbf{r}')$.

From (8.216) and (8.96) we can readily show that all linear functionals of a normal random process are normal. If we let $\mathbf{g} = \boldsymbol{\mathcal{H}}\mathbf{f}$, where $\boldsymbol{\mathcal{H}}$ is a linear CD mapping (see Sec. 7.3) defined by

$$g_m = \int_\infty d^q r \; h_m(\mathbf{r}) \, f(\mathbf{r}) \,, \qquad m = 1, ..., M \,, \tag{8.217}$$

then (8.96) becomes

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \exp(-2\pi i \mathbf{s}^\dagger \boldsymbol{\mathcal{H}} \overline{\mathbf{f}}) \exp(-2\pi^2 \mathbf{s}^\dagger \boldsymbol{\mathcal{H}} \boldsymbol{\mathcal{K}}_{\mathbf{f}} \boldsymbol{\mathcal{H}}^\dagger \mathbf{s}) \,. \tag{8.218}$$

By comparison with (8.196), we see that $\mathbf{g}$ is an MD random vector with mean $\boldsymbol{\mathcal{H}}\overline{\mathbf{f}}$ and covariance $\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_{\mathbf{f}}\boldsymbol{\mathcal{H}}^{\dagger}$.

Exactly the same conclusion holds when $\boldsymbol{\mathcal{H}}$ is an integral operator. Linear filtering of a normal random process yields another normal random process. Since normal processes are fully determined by their mean and autocovariance (or autocorrelation) function, the formulas given in Sec. 8.2.6 are all we need for a complete statistical description of the output of a linear filter *if* we know that the input is a normal process.

Multipoint densities and autocovariance functions One way of defining a normal random vector is to require that all of its marginals must be normal (Sec. 8.3.3). Similarly, a normal random process can be defined as one for which all univariate or multivariate marginals are normal. In that approach, a random process $f(\mathbf{r})$ is normal if all M-point PDFs, $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2), ..., f(\mathbf{r}_M)]$ for all M, are normal. We can use (8.218) to show that defining a normal random process by (8.216) is equivalent to requiring that all multipoint densities be normal. Evaluating the random process at the M points $\{\mathbf{r}_m, m = 1, ..., M\}$ is a CD mapping with

$$h_m(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_m)\,. \tag{8.219}$$

Thus $g_m = f(\mathbf{r}_m)$, and it follows at once from (8.218) that $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2), ..., f(\mathbf{r}_M)]$ is an MD normal density. An explicit form for this density can be stated most compactly by defining an $M \times 1$ vector $\mathbf{f}_M$ with m^{th} component given by $f(\mathbf{r}_m)$. For simplicity we assume that $f(\mathbf{r})$ is real. Then the M-point PDF is given by

$$\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2), ..., f(\mathbf{r}_M)] = \mathrm{pr}(\mathbf{f}_M)$$

$$= (2\pi)^{-\frac{1}{2}M} |\det \mathbf{K}_M|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\mathbf{f}_M - \overline{\mathbf{f}}_M)^t\, \mathbf{K}_M^{-1} (\mathbf{f}_M - \overline{\mathbf{f}}_M)\right] , \tag{8.220}$$

where $\overline{\mathbf{f}}_M$ is the $M \times 1$ mean vector, with components $\langle f(\mathbf{r}_m)\rangle$, and $\mathbf{K}_M$ is the $M \times M$ covariance matrix, with components given by

$$[\mathbf{K}_M]_{mn} = \langle [f(\mathbf{r}_m) - \langle f(\mathbf{r}_m)\rangle]\, [f(\mathbf{r}_n) - \langle f(\mathbf{r}_n)\rangle]\rangle\ . \tag{8.221}$$

Comparison with (8.98) shows that

$$[\mathbf{K}_M]_{mn} = K_{\mathbf{f}}(\mathbf{r}_m, \mathbf{r}_n)\,. \tag{8.222}$$

Thus the covariance *matrix* in an M-point PDF for a normal random process is fully determined by the autocovariance *function* of the process. Knowledge of this function and $\langle f(\mathbf{r})\rangle$ is therefore sufficient to specify all M-point densities and hence to fully characterize a normal process.

For completeness, we next show that (8.222) also follows from the transformation rule, $\mathbf{K}_{\mathbf{g}} = \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_{\mathbf{f}}\boldsymbol{\mathcal{H}}^{\dagger}$. With $\mathbf{K}_{\mathbf{g}} = \mathbf{K}_M$, and the kernel of $\boldsymbol{\mathcal{H}}$ as given by (8.219), we can write

$$[\mathbf{K}_M]_{mn} = \left[\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_{\mathbf{f}}\boldsymbol{\mathcal{H}}^{\dagger}\right]_{mn}$$

$$= \int_\infty d^q r \int_\infty d^q r'\, \delta(\mathbf{r} - \mathbf{r}_m) K_{\mathbf{f}}(\mathbf{r}, \mathbf{r}')\, \delta(\mathbf{r}' - \mathbf{r}_n) = K_{\mathbf{f}}(\mathbf{r}_m, \mathbf{r}_n)\,, \tag{8.223}$$

where the last step has used the sifting property of delta functions.

Ergodicity and stationarity Stationarity is defined for normal random processes just as for any other random process. A useful simplification, however, is that we do not have to distinguish wide-sense and narrow-sense stationarity in the normal case. Since the full statistics are inherent in the mean and autocovariance function, wide-sense stationarity (stationary mean and autocovariance) implies narrow-sense or strict stationarity (Papoulis, 1965).

For stationary Gaussian random processes, a straightforward criterion for ergodicity can be stated. Cornfield *et al.* (1982) show that such a process is ergodic if and only if its power spectral density is continuous. From (3.107) and the Wiener-Khinchin theorem (8.133), an equivalent statement is that a stationary Gaussian random process is ergodic if and only if its autocorrelation function vanishes at infinity. Since many physical processes are Gaussian as a result of the central-limit theorem, we can quite often invoke ergodicity on the basis of this theorem.

8.3.6 Complex Gaussian random fields

It is often useful to describe a wave by its complex amplitude. If the wave is regarded as random, perhaps because it has been scattered from a random object, then the wave amplitude $u(\mathbf{r}_1)$ at any point $\mathbf{r}_1$ is a complex random variable. Similarly, the set of amplitudes at K different points, $\{u(\mathbf{r}_k), k = 1, ..., K\}$, is a KD complex vector, and $u(\mathbf{r})$ itself is a complex random process. Moreover, a wave amplitude is usually computed as a diffraction integral or some other linear superposition. If different elements of this superposition are linearly independent random variables, then the central-limit theorem will lead to normal distributions, so we often encounter complex Gaussian random fields.

In one sense, there is nothing new about complex Gaussian random fields; we can describe them with the tools already developed for real Gaussian fields just by considering the real and imaginary parts separately. For example, a $K \times 1$ complex vector can also be written as a $2K \times 1$ real vector, where the first K components are the real parts and the second K are the imaginary parts. The covariance matrix in the first case is a $K \times K$ Hermitian matrix with complex off-diagonal elements, and in the second case it is a $2K \times 2K$ real, symmetric matrix.

Random phase If the complex variables result from random waves, the physics of wave propagation may allow us to impose some additional restrictions, thereby simplifying the mathematics. The phase of a wave relates to the total optical pathlength from a radiation source to the point at which the phase is measured. The natural unit of this pathlength is the wavelength, and typically the paths are very long compared to a wavelength. That means that if we alter the pathlength by a small fraction in absolute terms, it may nevertheless change by several wavelengths, and each change of one wavelength alters the phase by 2π. Now, the pathlength (in units of wavelength) may be random for many reasons: we may consider an ensemble of objects with different positions and different rough surfaces, or we may interpose random phase-altering elements such as diffuse reflectors or ground-glass screens, or we may consider a broad spectrum of wavelengths. The result is that it is frequently an excellent approximation to assume that the phase is completely random.

To state this approximation more mathematically, we denote the wave amplitude (at some unspecified point) by $u = Ae^{i\phi} = x + iy$, where $x = \mathrm{Re}(u)$ and

$y = \mathrm{Im}(u)$ and A is a real number. We do not need to consider phase angles ϕ outside the range $[0, 2\pi)$ since $e^{i\phi}$ is periodic. The phase randomness implies that the PDF on ϕ is constant in this range. The constant can be fixed since the PDF must be normalized to unity, and we can write

$$\mathrm{pr}(\phi) = \frac{1}{2\pi}\,, \qquad 0 \le \phi < 2\pi\,. \tag{8.224}$$

We assume that this PDF on ϕ is valid for all A, so $\mathrm{pr}(\phi|A) = \mathrm{pr}(\phi)$, and ϕ and A are statistically independent.

We can use this density to deduce some important properties of u even without specifying the statistics of A. Since the real and imaginary parts of u are given by

$$x = A\cos\phi\,, \qquad y = A\sin\phi\,, \tag{8.225}$$

we see that (8.224) implies

$$\langle x\rangle = \langle A\cos\phi\rangle = 0, \qquad \langle y\rangle = \langle A\sin\phi\rangle = 0\,. \tag{8.226}$$

Thus x and y are both zero-mean, and hence so is the complex u.

The variances of x and y must be equal since

$$\left\langle x^2\right\rangle = \left\langle A^2\cos^2\phi\right\rangle = \tfrac{1}{2}\left\langle A^2\right\rangle\,, \qquad \left\langle y^2\right\rangle = \left\langle A^2\sin^2\phi\right\rangle = \tfrac{1}{2}\left\langle A^2\right\rangle\,. \tag{8.227}$$

The marginal PDFs on x and y must also be the same, regardless of the PDF of A, since $\sin\phi$ and $\cos\phi$ have the same PDFs if ϕ is uniform. (As an exercise, the reader can determine what this PDF is.) Moreover, x and y are uncorrelated since

$$\langle xy\rangle = \left\langle A^2\cos\phi\sin\phi\right\rangle = 0\,. \tag{8.228}$$

We can summarize the last two equations in complex form by writing

$$\left\langle u^2\right\rangle = \left\langle u^{*2}\right\rangle = 0\,, \qquad \langle uu^*\rangle = \langle u^*u\rangle = \left\langle A^2\right\rangle \ne 0\,. \tag{8.229}$$

Invocation of the central-limit theorem If we now assume that the wave amplitude at any point is the sum of contributions from many independent sources (perhaps points on an illuminated rough surface), then the real and imaginary parts are normal by the central-limit theorem. That means that x and y are not only uncorrelated but also statistically independent; we say that x and y are i.i.d. (independently and identically distributed). Their joint density is given by

$$\mathrm{pr}(x, y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)\,, \tag{8.230}$$

where σ^2 is the common variance of x and y. Contours of constant PDF in the x-y plane are circles (see Fig. 8.3), so u is referred to as a *circular Gaussian* random variable.

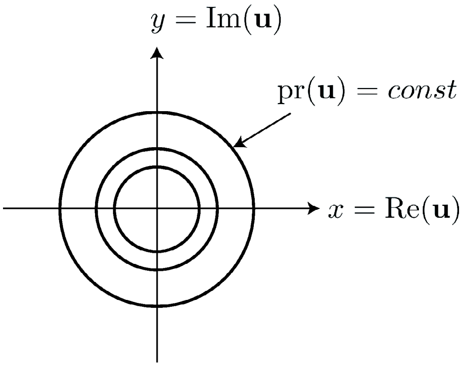

$y = \mathrm{Im}(\mathbf{u})$
$\mathrm{pr}(\mathbf{u}) = const$
$x = \mathrm{Re}(\mathbf{u})$

Fig. 8.3 Surfaces of constant probability density for a circular Gaussian random variable.

Other useful PDFs Since $A = \sqrt{x^2+y^2}$ and $\phi = \tan^{-1}(y/x)$, we can convert $\mathrm{pr}(x,y)$ in (8.230) to $\mathrm{pr}(A,\phi)$ by means of (C.104). The result is the *Rayleigh distribution*, given in (C.140) as

$$\mathrm{pr}(A,\phi) = \frac{A}{2\pi\sigma^2}\exp\left(-\frac{A^2}{2\sigma^2}\right) . \tag{8.231}$$

We shall see in Chap. 11 that the irradiance $I = |u|^2$ plays a key role in photodetection and photon counting. If $\mathrm{pr}(x,y)$ is given by (8.230), the PDF on I and ϕ is

$$\mathrm{pr}(I,\phi) = \frac{1}{2\pi\overline{I}}\exp\left(\frac{-I}{\overline{I}}\right) , \tag{8.232}$$

where $\overline{I} = 2\sigma^2$. The PDF on I alone, obtained by omitting the 2π in (8.232), is a chi-squared PDF with two degrees of freedom (see Sec. C.5.5). In general, a chi-squared random variable with N degrees of freedom is the sum of the squares of N i.i.d. normal random variables; here $N = 2$ since $I = x^2 + y^2$.

Two-point densities for circular Gaussians Next we examine two-point PDFs involving a complex circular Gaussian random process $u(\mathbf{r})$ at points $\mathbf{r} = \mathbf{r}_1$ and $\mathbf{r} = \mathbf{r}_2$. For notational simplicity, we write $u(\mathbf{r}_1) = u_1 = x_1 + iy_1 = A_1\exp(i\phi_1)$, and similarly for $u(\mathbf{r}_2)$. One way we could specify the two-point density would be to construct the real 4D vector $\mathbf{U} = (x_1, x_2, y_1, y_2)^t$ and give the 4D PDF for it. If $u(\mathbf{r})$ is to be circular Gaussian, this PDF has to satisfy some constraints. For one thing, if we want u_1 and u_2 to be individual circular Gaussians, the marginals on (x_1, y_1) and (x_2, y_2) must both satisfy (8.230), possibly with different variances. In addition, the joint density on all four variables must be consistent with the autocovariance function of the process,

$$K_{\mathbf{u}}(\mathbf{r}_1,\mathbf{r}_2) = \langle u_1 u_2^*\rangle \equiv k = k' + ik'' . \tag{8.233}$$

These conditions lead to

$$\langle x_1x_2\rangle = \langle y_1y_2\rangle = \tfrac{1}{2}k' , \qquad -\langle x_1y_2\rangle = \langle y_1x_2\rangle = \tfrac{1}{2}k'' \qquad \langle x_1y_1\rangle = \langle x_2y_2\rangle = 0 . \tag{8.234}$$

All of these conditions are satisfied if $\mathbf{U} \sim \mathcal{N}_4(\mathbf{0}, \mathbf{K_U})$, where

$$\mathbf{K_U} = \begin{bmatrix} \sigma_1^2 & \frac{1}{2}k' & 0 & -\frac{1}{2}k'' \\ \frac{1}{2}k' & \sigma_2^2 & \frac{1}{2}k'' & 0 \\ 0 & \frac{1}{2}k'' & \sigma_1^2 & \frac{1}{2}k' \\ -\frac{1}{2}k'' & 0 & \frac{1}{2}k' & \sigma_2^2 \end{bmatrix} . \tag{8.235}$$

The redundancy in the elements of this matrix should be noted. A general 4×4 covariance matrix would have 10 independent elements, but only four real numbers ($\sigma_1^2, \sigma_2^2, k'$ and k'') are required to specify $\mathbf{K}_U$. This redundancy is required in order to represent a circular Gaussian as opposed to a more general complex Gaussian random vector.

Two-dimensional formulation To go from the covariance in (8.235) to the PDF for $\mathbf{U}$ requires inverting $\mathbf{K_U}$ and computing the quadratic form $\mathbf{U}^t\mathbf{K_U}^{-1}\mathbf{U}$. The algebra is not terrible, but a simpler approach, and one that extends more readily to higher dimensions, is to use a 2D complex vector rather than a 4D real one. If we define a 2D vector $\mathbf{u}$ with complex components u_1 and u_2, its covariance matrix is

$$\mathbf{K_u} = \begin{bmatrix} 2\sigma_1^2 & k \\ k^* & 2\sigma_2^2 \end{bmatrix} . \tag{8.236}$$

The inverse covariance, which is what we need in the PDF, is given by

$$\mathbf{K_u}^{-1} = \frac{1}{4\sigma_1^2\sigma_2^2 - |k|^2} \begin{bmatrix} 2\sigma_2^2 & -k \\ -k^* & 2\sigma_1^2 \end{bmatrix} . \tag{8.237}$$

The quadratic form in the PDF is thus

$$\mathbf{u}^\dagger \mathbf{K_u}^{-1} \mathbf{u} = \frac{2\sigma_2^2|u_1|^2 + 2\sigma_1^2|u_2|^2 - ku_1^*u_2 - k^*u_2^*u_1}{4\sigma_1^2\sigma_2^2 - |k|^2} , \tag{8.238}$$

and the corresponding PDF is given by (Neeser and Massey, 1993; Mandel and Wolf, 1995)

$$\mathrm{pr}(\mathbf{u}) = \frac{1}{\pi^2 \det(\mathbf{K_u})} \exp\left(-\mathbf{u}^\dagger \mathbf{K_u}^{-1} \mathbf{u}\right) . \tag{8.239}$$

The reader might have expected a factor of $\frac{1}{2}$ in the exponent and a different normalizing factor [*cf.* (8.185)], but (8.239) is correct as written. One way to make it plausible is to assume there is no correlation, so $k = 0$, so that (8.237) becomes

$$\mathbf{K_u}^{-1} = \begin{bmatrix} \frac{1}{2\sigma_1^2} & 0 \\ 0 & \frac{1}{2\sigma_2^2} \end{bmatrix} . \tag{8.240}$$

Hence, (8.239) becomes

$$\mathrm{pr}(\mathbf{u}) = \frac{1}{4\pi^2\sigma_1^2\sigma_2^2} \exp\left[-\frac{x_1^2+y_1^2}{2\sigma_1^2} - \frac{x_2^2+y_2^2}{2\sigma_2^2}\right] , \tag{8.241}$$

which is just what one would get with the 4D real formulation, using (8.235) with $k = 0$ and (8.185). The reader may check that the 2D complex and 4D real formulations also agree when $k \neq 0$. (The 4D determinant must be evaluated by minors.)

Complex Gaussian vectors Most authors use the $2N$D real formulation to deal with ND complex random vectors, but there is a significant literature on the complex formulation. The classic text by Doob (1953) discusses the problem, and Wooding (1956) first derived a form like (8.239).

Later authors, however, recognized some surprising features of the complex case (Reed, 1962; Goodman, 1963; Neeser and Massey, 1993). For example, we must revisit the familiar statement that the PDF for a Gaussian random vector is fully determined by its covariance matrix. For a complex random vector, the covariance is defined by $\mathbf{K_u} = \langle(\mathbf{u} - \overline{\mathbf{u}})(\mathbf{u} - \overline{\mathbf{u}})^\dagger\rangle$, but the most general PDF for a Gaussian random vector also involves the *pseudocovariance* $\langle(\mathbf{u} - \overline{\mathbf{u}})(\mathbf{u} - \overline{\mathbf{u}})^t\rangle$, with a transpose in place of the adjoint.

As defined by Neeser and Massey (1993), a complex random vector is said to be *proper* if its pseudocovariance vanishes identically. Any subvector of a proper random vector is proper, but two individually proper random vectors are not necessarily jointly proper. These authors also show that any linear or affine transformation of a proper random vector is another proper random vector, and that a real random vector can be proper if and only if it is a constant.

The condition that the pseudocovariance of a complex vector vanish can be restated in terms its real and imaginary components. If we write $\mathbf{u} = \mathbf{x} + i\mathbf{y}$, then $\mathbf{u}$ is proper if and only if

$$\langle(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^t\rangle = \langle(\mathbf{y}-\overline{\mathbf{y}})(\mathbf{y}-\overline{\mathbf{y}})^t\rangle \quad \text{and} \quad \langle(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{y}-\overline{\mathbf{y}})^t\rangle = -\langle(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{y}-\overline{\mathbf{y}})^t\rangle^t . \tag{8.242}$$

Thus $\mathbf{x}$ and $\mathbf{y}$ must have identical autocovariance matrices, and their cross-covariance matrix must be skew-symmetric.

For optical applications, we are often interested in zero-mean proper Gaussian random vectors and processes, for which the term *circular Gaussian* is commonly used. To be explicit, an ND complex vector $\mathbf{u}$ will be said to obey a circular Gaussian law if all marginals are normal, all components have zero mean and the conditions in (8.242) hold; these conditions can be stated in complex form as

$$\langle u_n u_m\rangle = \langle u_n^* u_m^*\rangle = 0\,, \qquad 1 \le n,\, m \le N \tag{8.243}$$

and

$$\langle u_n u_m^*\rangle = \langle u_m u_n^*\rangle^* = K_{nm}\,. \tag{8.244}$$

The intuition behind (8.243) is that u_n can be written as $|u_n| \exp(i\phi_n)$, where ϕ_n is uniformly distributed over $(0, 2\pi)$ but possibly correlated with ϕ_m for $n \neq m$. The expectation $\langle u_n u_m\rangle$ is zero because $\exp[i(\phi_n + \phi_m)]$ takes any value on the unit circle with equal probability. One can think of choosing a ϕ_n from the conditional density $\mathrm{pr}(\phi_n|\phi_m)$ and then choosing ϕ_m from the uniform density; no matter what ϕ_n is chosen in the first step, the second choice means that $\phi_n + \phi_m$ (modulo 2π) is equally likely to be anywhere in $(0, 2\pi)$. On the other hand, $\langle u_n u_m^*\rangle$ depends on $\exp[i(\phi_n - \phi_m)]$, and this average is not zero if ϕ_n and ϕ_m tend to fluctuate together; the second choice tends to undo the first.

The PDF of an ND circular Gaussian random vector is a generalization of (8.239):

$$\mathrm{pr}(\mathbf{u}) = \frac{1}{\pi^N \det(\mathbf{K_u})} \exp\left(-\mathbf{u}^\dagger \mathbf{K}_\mathbf{u}^{-1} \mathbf{u}\right) . \tag{8.245}$$

Thus the only change in going from 2D to ND is the power of π. It is proven in Bellman (1995) that this density is properly normalized, and the reader can check it by considering the basis in which $\mathbf{K_u}$ is diagonal.

The characteristic function for complex random vectors is defined in (8.33); for an ND circular Gaussian it is given by

$$\psi_\mathbf{u}(\boldsymbol{\xi}) = \exp(-\pi^2 \boldsymbol{\xi}^\dagger \mathbf{K_u} \boldsymbol{\xi}) , \tag{8.246}$$

where $\boldsymbol{\xi}$ is an ND complex vector. Note the absence of a factor of 2 in the exponent when compared to the corresponding expression (8.196) for a real Gaussian random vector.

Moments The characteristic function can be used to derive all moments of a random vector. For complex random vectors, the rules for complex differentiation given in Sec. A.9.5 must be used. The reader may use these rules to verify that (8.246) is consistent with the second moments stated in (8.243) and (8.244).

Higher moments are also of interest in many problems. For circular Gaussians, all odd moments vanish, as do all even moments where the number of factors without the complex conjugate is not equal to the number with the conjugate. All other even moments can be expressed in terms of components of the covariance matrix via the *complex Gaussian moment theorem*, first derived by Reed (1962) and discussed by Goodman (1985) in terms of real components and by Osche (2002) in complex form. Osche's statement of the theorem is

$$\langle u_{n_1}\, u_{n_2} \,\cdots\, u_{n_t}\, u^*_{m_1}\, u^*_{m_2} \,\cdots\, u^*_{m_t} \rangle = \sum_\pi \langle u_{n_1}\, u^*_{m_{\pi(1)}} \rangle \langle u_{n_2}\, u^*_{m_{\pi(2)}} \rangle \,\cdots\, \langle u_{n_t}\, u^*_{m_{\pi(t)}} \rangle , \tag{8.247}$$

where $\pi(\,\cdot\,)$ is a permutation of the set of integers $\{1,\, 2,\, \cdots,\, t\}$, and the sum is over all possible permutations. Some useful special cases are:

$$\langle |u_i|^{2n} \rangle = n!\, \langle |u_i|^2 \rangle^n = n!\, \sigma_i^{2n} \, ; \tag{8.248}$$

$$\langle (u_i u_j^*)^n \rangle = n!\, \langle u_i u_j^* \rangle^n = n!\, K_{ij}^n \, ; \tag{8.249}$$

$$\langle u_i u_j u_k^* u_\ell^* \rangle = \langle u_i u_k^* \rangle \langle u_j u_\ell^* \rangle + \langle u_j u_k^* \rangle \langle u_i u_\ell^* \rangle = K_{ik} K_{j\ell} + K_{jk} K_{i\ell} \, . \tag{8.250}$$

This latter equation should be compared to the corresponding real result in (8.197); the complex expression has a sum of two covariances while the real expression has three. We see that $\left\langle |u_i|^4 \right\rangle = 2\sigma_i^4$, but for a real, zero-mean, Gaussian random variable, $\left\langle g_i^4 \right\rangle = 3\sigma_i^4$. The reader can verify this result by writing $u_i = x_i + iy_i$ and using the real Gaussian moment theorem.

Circular Gaussian random processes A complex random process $u(\mathbf{r})$ will be said to be circular Gaussian if all N-point PDFs are multivariate circular Gaussian random vectors. We can specify this process, as in Sec. 8.3.5, by its characteristic *functional*, given by [*cf.* (8.216)]

$$\Psi_\mathbf{u}(\mathbf{s}) = \exp(-\pi^2 \mathbf{s}^\dagger \boldsymbol{\mathcal{K}}_\mathbf{u} \mathbf{s}) , \tag{8.251}$$

where $\mathbf{s}$ is a square-integrable function and $\boldsymbol{\mathcal{K}}_{\mathbf{f}}$ is the autocovariance operator, *i.e.*, the integral operator with kernel $K_{\mathbf{u}}(\mathbf{r},\mathbf{r}') = \langle u(\mathbf{r})\, u^*(\mathbf{r}')\rangle$. We shall make good use of (8.251) in Chap. 18 when we discuss speckle.

8.4 STOCHASTIC MODELS FOR OBJECTS

We argued in Chap. 7 that an object was best described by a function $f(\mathbf{r})$ (where $\mathbf{r}$ is usually a position vector); now we shall regard this function as a sample function of a random process. The random process is the collection of all possible objects of a given category that might be presented to the imaging system. For example, in computed tomography of the brain, a particular object $f(\mathbf{r})$ is one patient's brain at the time of one imaging procedure, but we can imagine an infinite ensemble of brains from which this one object is drawn. Ideally we would like to specify the full, infinite-dimensional, probability density function (PDF) of the process. As we shall see in Sec. 8.4.1, however, a full PDF is seldom possible, even in principle, and we must make do with less complete models.

The literature on stochastic models in image science is rich and varied, but often the distinction between an object model and an image model is not clear. Many papers claim to address the statistics of images but leave out any consideration of measurement noise or system blur. Moreover, these papers often treat the image as a function of continuous spatial coordinates rather than as a discrete array. Thus they really apply more to objects than to real-world images. On the other hand, if we want to verify our theories by measurements, all we have access to is images, and there is a gap in the current literature on how one can verify stochastic models of objects from observations on noisy, blurred, discrete images.

Another confusing aspect of much of the literature has to do with the meaning of probability. First, there is an unfortunate emphasis on ergodic models where it is assumed, often tacitly, that probabilistic statements can be made for a single object or image. Thus a gray-level histogram of a single image is treated as a probability distribution for pixel values. At best the histogram is an estimate of the probability law for an ensemble of similar images, and then only if ergodicity and hence stationarity are assumed. Except for relatively contrived situations, stationarity is unlikely to hold over the full expanse of an object or image (though local stationarity may be more defensible).

Closely associated with the emphasis on stationarity is the use of loosely defined Fourier measures called *power spectra*. Often this term refers to nothing more than the square modulus of the Fourier transform of a single image. With an assumption of ergodicity this quantity is an estimate of the power spectral density, defined in Sec. 8.2.5 as the Fourier transform of the statistical autocorrelation function. We know from Fig. 8.1, however, that the estimate is poor, and in any case the implicit statistical ensemble is seldom specified, and the underlying stationarity assumption is almost never justified.

Another issue is the conflict between Bayesian and frequentist interpretations of probability, introduced in the Prologue. For many purposes, we want models that emulate reality, in the sense that the model predictions can be verified in principle by measurements on real objects, so we are using a frequentist interpretation of probability. Bayesian interpretations of probability are often useful, however, especially in drawing inferences from images when we have some degree of prior belief about the structure of the object but the frequentist information is incomplete (as it always is). The use of Bayesian priors will be explored further in Chaps. 13 and 15, but the emphasis in this section is descriptive: What can we say about collections of real objects?

In practice, even the very concept of a *real object* must often be expanded. Computer simulations are becoming ever more realistic and ever more essential in image science, and we do not rule out collections of simulations as the ensemble of objects for which we seek a stochastic model.

To state clearly the focus of this section, then, we are considering an ensemble interpretation of probability as applied to objects regarded as sample functions of a random process. The sample function can, in principle, be an actual object $f(\mathbf{r})$, but in practice it may be some approximate representation $f_a(\mathbf{r})$ as introduced in Sec. 7.1.3, and the object can be simulated rather than real.

We begin in Sec. 8.4.1 with a general discussion of just what we mean by the probability density function for an object class and how we might approach the problem experimentally. Included in this section is an introduction to the important concept of independent components.

In Sec. 8.4.2 we revisit the discussion from Sec. 8.2.2 on multipoint densities, but now specifically for objects. Again the focus is on experimental determination of stochastic models.

In Sec. 8.4.3 we do what all statisticians do when problems get difficult: we assume normality. Some implications of the central-limit theorem are discussed, and Gaussian mixture models are introduced. Surprisingly, Gaussian mixture models turn out to account for the highly non-Gaussian character of many filtered images.

In Sec. 8.4.4 we turn to the widely studied but loosely defined topic of texture. For purposes of this section, a texture is regarded as any random field with some degree of stationarity. We discuss here ways of synthesizing sample textures as well as mathematical models for the PDFs.

Sec. 8.4.5 is prelude to the discussion of signal detection in Chap. 13. We make a distinction between signals and backgrounds, and we look at how various assumptions about the signal affect the overall object PDF.

8.4.1 Probability density functions in Hilbert space

To develop a Hilbert-space PDF for objects, we assume that a function $f(\mathbf{r})$ representing a particular object is square-integrable and therefore corresponds to a vector $\mathbf{f}$ in $\mathbb{L}_2(\mathbf{S}_f)$, where $\mathbf{S}_f$ is a support region that will cover all object functions under consideration. Then $\mathbf{f}$ can be expanded as in (8.76):

$$\mathbf{f} = \sum_{n=1}^{\infty} \alpha_n \boldsymbol{\psi}_n \, , \tag{8.252}$$

where the set $\{\boldsymbol{\psi}_n\}$ is some convenient basis for $\mathbb{L}_2(\mathbf{S}_f)$. The coefficients $\{\alpha_n\}$ are the components of $\mathbf{f}$ in this basis. If the basis is orthonormal, the infinite-dimensional vector of coefficients, denoted $\boldsymbol{\alpha}$, is a unitary transformation of $\mathbf{f}$.

Intuitively, $\mathbf{f}$ corresponds to a single point in the space (or a vector from the origin to the point), and the density $\mathrm{pr}(\mathbf{f})$ is a measure of how these points cluster in the Hilbert space. The density $\mathrm{pr}(\boldsymbol{\alpha})$ describes this same clustering in terms of specific basis vectors $\boldsymbol{\psi}_n$.

A graphical depiction of this clustering is shown in Fig. 8.4. The two axes shown can be construed as any two components $\{\alpha_n, \alpha_m\}$ out of the infinite set.

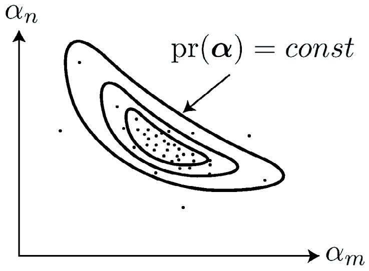


Fig. 8.4 Graphical depiction of the clustering of an object PDF. Two axes out of an infinite-dimensional Hilbert space are shown, and each point corresponds to a different object.

Subspaces We can never hope to know the full PDF in an infinite-dimensional space (and we wouldn't know what to do with it if we had it), but our ultimate goal is always to obtain a PDF $\mathrm{pr}(\mathbf{g})$ for images (see Sec. 8.5). Since the data are insensitive to null functions of the imaging operator $\boldsymbol{\mathcal{H}}$, and all real measurement operators have finite rank R, we can always get by with a finite-dimensional subspace of the object space $\mathbb{U}$. As we know from Sec. 7.4.3, we can use the singular vectors of $\boldsymbol{\mathcal{H}}$ as the expansion functions and truncate the expansion at $n = R$; this truncation produces no error in the data and hence no error in $\mathrm{pr}(\mathbf{g})$.

Another way to restrict the dimensionality is to construct an approximate representation of $\mathbf{f}$, just as we did in Chap. 7, and then consider the PDF of the approximate vector $\mathbf{f}_a$. This procedure can lead to an error in $\mathrm{pr}(\mathbf{g})$, but it will be small if the image error defined in Sec. 7.4.3 is small for all objects in the ensemble. In fact, the image error will be zero if we use natural pixels as the expansion functions (see Sec. 7.4.3).

Experimental determination of the object density We can imagine obtaining information about the object density by examining a large number of typical object functions. There are several ways we could know the object function. For example, we might use a computer program that can simulate sample functions $f(\mathbf{r})$, and for each sample function we could obtain components α_n by computing scalar products with the corresponding basis functions $\psi_n(\mathbf{r})$. (In fact, if a set of components is chosen in advance, the computer program could advantageously generate the sample functions in this basis in the first place.)

Alternatively, we may want to construct a stochastic model useful for one particular imaging system, say a relatively low-resolution, noisy one, but we might have available images from another system with better resolution and less noise. We could then treat the *images* from the better system as good representations of *objects* for the poorer system.

Finally, we might have some physical model, known as a phantom in the medical-imaging literature. If the phantom can be reconfigured into different objects by moving components around in a controllable fashion, it can generate a set of known sample objects.

With any of these sources of sample objects, a histogram estimate of, say, $\mathrm{pr}(\alpha_n, \alpha_m)$ could be obtained by a frequentist interpretation of the PDF. By a multivariate generalization of (C.21), we can write[8]

$$\mathrm{pr}_{(\alpha_n,\alpha_m)}(\alpha_{nk},\alpha_{mk})$$

$$\equiv \lim_{\Delta\alpha\to 0}\frac{1}{(\Delta\alpha)^2}\,\mathrm{Pr}(\alpha_{nk}-\tfrac{1}{2}\Delta\alpha \le \alpha_n < \alpha_{nk}+\tfrac{1}{2}\Delta\alpha, \alpha_{mk}-\tfrac{1}{2}\Delta\alpha \le \alpha_m < \alpha_{mk}+\tfrac{1}{2}\Delta\alpha)\,. \tag{8.253}$$

The histogram estimate is obtained by considering finite bins of width $\Delta\alpha$ (hence omitting the limit) and approximating the probabilities on the right with observed frequencies of occurrence in a finite number of sample objects. Thus we approximate the density as

$$\widehat{\mathrm{pr}}_{(\alpha_n,\alpha_m)}(\alpha_{nk},\alpha_{mk}) \equiv \frac{1}{(\Delta\alpha)^2}\,\frac{J(\alpha_{nk},\alpha_{mk})}{J}\,, \tag{8.254}$$

where $J(\alpha_{nk},\alpha_{mk})$ is the number of times (out of J sample objects) that the computed value of (α_n,α_m) falls in a square of size $(\Delta\alpha)^2$ centered on point $(\alpha_{nk},\alpha_{mk})$. This estimate can, in principle, be extended to an arbitrary number of dimensions.

The problem with this scenario is that the required number of samples may be impractical. As a numerical example, suppose the objects can be adequately specified by 10^4 terms in (8.252), so we are seeking to construct a histogram approximation to a PDF in a ten-thousand-dimensional space. If we choose to use just 10 bins along each axis in the space, then there are $10^{10,000}$ total bins to fill. This is an immense[9] number, and there is no hope of filling the bins with experimental samples. Even with a drastically truncated set of components, $\mathrm{pr}(\boldsymbol{\alpha})$ cannot be interpreted in frequentist terms.

Independent components The number of samples required for a histogram estimate would be much smaller if the components were statistically independent. In that case, for an ND representation, we would have

$$\mathrm{pr}(\boldsymbol{\alpha}) = \prod_{n=1}^{N}\mathrm{pr}(\alpha_n)\,, \tag{8.255}$$

so we would need only a set of N univariate densities instead of an N-dimensional multivariate one.

In contrast to $\mathrm{pr}(\boldsymbol{\alpha})$, the univariate density $\mathrm{pr}(\alpha_n)$ does admit of a frequentist interpretation and a histogram estimate. Suppose, as above, that we have some source of object functions $f(\mathbf{r})$, perhaps a computer simulation code. For each sample function we can evaluate α_n by the usual scalar product, and the histogram estimate of $\mathrm{pr}(\alpha_n)$ is defined by [*cf.* (8.254)]

$$\widehat{\mathrm{pr}}_{\alpha_n}(\alpha_{nk}) = \frac{1}{\Delta\alpha}\,\frac{J_{nk}}{J}\,, \tag{8.256}$$

---

[8] Recall our notational convention that subscripts on PDFs are deleted where they are redundant with the argument. Thus $\mathrm{pr}(x)$ and $\mathrm{pr}_x(x)$ mean the same thing but the subscript is reinstated on $\mathrm{pr}_x(x_0)$, which means $\mathrm{pr}_x(x)$ evaluated at $x = x_0$.

[9] We use the term *immense* here in its literal sense: incapable of mensuration, immeasurable. Certainly any number exceeding the number of atoms in the universe (of order 10^{80}) qualifies as immense.

where α_{nk} is the specific value of α_n centered on the k^{th} bin, and J_{nk} is the number of times α_n falls in that bin.

The number J_{nk} is a random variable; if the experiment is repeated many times with different sample objects, J_{nk} will be binomially distributed, and the full set of J_{nk} values will be multinomially distributed (see Secs. C.6.1. and 11.2.1). The mean value of J_{nk} will be J times the probability that α_n falls in bin k, or

$$\langle J_{nk} \rangle \approx J \, \mathrm{pr}_{\alpha_n}(\alpha_{nk}) \, \Delta\alpha \, . \tag{8.257}$$

If the number of bins is large, the probability that α_n will fall in one particular bin is small, and any reasonable experiment will use a large value for J, so we are dealing with rare events (see Sec. 11.1.2) where the binomial law on J_{nk} is well approximated by a Poisson.

As a practical example, suppose we want to construct a 100-bin histogram. By the Poisson statistics, a relative error (standard deviation divided by mean) of 10% in the value estimated for the k^{th} bin requires $\langle J_{nk} \rangle = 100$, and a relative error of 1% requires $\langle J_{nk} \rangle = 10^4$. To relate these numbers to the required number of images, we must make some assumptions about the underlying distribution of α_n. If we assume that $\mathrm{pr}(\alpha_n)$ is relatively flat over the range from 0 to α_{max}, then each $\langle J_{nk} \rangle$ is approximately J divided by the number of bins, or $0.01J$ in our example. Thus we require $J = 10^4$ for 10% accuracy and 10^6 for 1% accuracy in a 100-bin histogram. These numbers are large but not immense; they are well within the capabilities of modern computers if the sample objects are simulated. Moreover, each simulated object can be used to evaluate each α_n, so we get the full multivariate PDF for this amount of simulation effort, but only if the components are independent.

Finding the independent components One approach to finding approximately independent components is the Karhunen-Loève (KL) expansion, introduced in Sec. 7.2.4. In Sec. 8.2.7 we showed that the KL expansion yields uncorrelated coefficients, and if we can argue that the process is Gaussian (see Sec. 8.4.3), then uncorrelated implies independent.

To use this argument, we must know the KL expansion. For stationary random processes, as discussed in Sec. 8.2.4, KL expansion is Fourier analysis, but with nonstationary models it can be difficult to determine the autocorrelation function, much less to diagonalize it and find the KL basis. As we shall see in Sec. 8.4.5, some authors argue that wavelet coefficients are approximately uncorrelated for natural scenes, so a wavelet transformation is approximately a KL transformation. Even when this argument can be justified, however, it is still necessary to show that the wavelet coefficients are Gaussian random variables if we want to use (8.255), and we shall present an argument in Sec. 8.4.3 showing why this is *not* the case for a wide class of natural scenes.

When the process is not Gaussian or when we do not know the KL expansion, it may nevertheless be possible to find a transformation that makes the expansion coefficients approximately independent. To make this statement more precise, we need some definition of degree of dependence.

One way to define degree of dependence is in terms of the distance, in some sense, between the multivariate density and the product of its marginals. One distance measure used for this purpose is the *Kullback-Leibler distance*, known also as the *cross-entropy* or *mutual information.* If we consider an $N \times 1$ vector $\boldsymbol{\beta}$ with density $\mathrm{pr}(\boldsymbol{\beta})$, the Kullback-Leibler distance between this density and the product of its marginals is defined by (Comon, 1994)

$$I(\boldsymbol{\beta}) = \int_\infty d^N\beta \; \mathrm{pr}(\boldsymbol{\beta}) \ln \left\{ \frac{\mathrm{pr}(\boldsymbol{\beta})}{\prod_{n=1}^N \mathrm{pr}(\beta_n)} \right\} . \tag{8.258}$$

Note that $I(\boldsymbol{\beta})$ is not a true distance, as defined in Sec. 1.1.2, since it is not symmetric in interchange of $\mathrm{pr}(\boldsymbol{\beta})$ and $\prod_{n=1}^N \mathrm{pr}(\beta_n)$. It does, however, vanish when these two densities are equal, since the argument of the logarithm is unity in that case, and it follows from the convexity of the logarithm that $I(\boldsymbol{\beta}) \geq 0$ (Kendall and Stuart, 1979). Thus independent components can be sought by attempting to find a basis that minimizes $I(\boldsymbol{\beta})$.

Many other measures of degree of dependence are discussed by Comon (1994). In particular, he uses an Edgeworth approximation to argue that independent components will have marginals with large kurtoses, as defined in (C.41). He therefore suggests maximizing the sum of the squares of the marginal kurtoses as as a way of finding approximately independent components. We refer the reader to Comon (1994) for a full justification of this approach.

Independent components analysis A structured approach to minimizing some measure of statistical dependence is *independent components analysis* or ICA. ICA is a refinement of *principal components analysis* or PCA, which we shall discuss first.

Though the terms PCA and KL are often used interchangeably in the literature, we make the distinction that PCA is diagonalization of the sample covariance matrix and KL is based on the ensemble covariance. Thus PCA approaches KL analysis as the number of samples goes to infinity.

Suppose we are given J samples of a random vector $\boldsymbol{\alpha}$, denoting the j^{th} sample by $\boldsymbol{\alpha}^{(j)}$. The sample covariance matrix $\widehat{\mathbf{K}}_{\boldsymbol{\alpha}}$ is defined by

$$\widehat{\mathbf{K}}_{\boldsymbol{\alpha}} = \frac{1}{J} \sum_{j=1}^{J} \left[\Delta\boldsymbol{\alpha}^{(j)}\right] \left[\Delta\boldsymbol{\alpha}^{(j)}\right]^\dagger , \tag{8.259}$$

where $\Delta\boldsymbol{\alpha}^{(j)}$ is $\boldsymbol{\alpha}^{(j)}$ minus the sample mean. PCA seeks to find a matrix $\mathbf{M}$ such that the transformed sample vectors,

$$\boldsymbol{\beta}^{(j)} = \mathbf{M}\boldsymbol{\alpha}^{(j)} , \tag{8.260}$$

are uncorrelated and hence the transformed sample covariance matrix $\widehat{\mathbf{K}}_{\boldsymbol{\beta}}$ is diagonal. By retracing the discussion in Sec. 8.1.6 but with $\widehat{\mathbf{K}}$ in place of $\mathbf{K}$, we can see that this diagonalization is accomplished by using the eigenvectors of $\widehat{\mathbf{K}}_{\boldsymbol{\alpha}}$ as the columns of $\mathbf{M}$.

ICA also uses a transformation of the form (8.260), but now the goal is to minimize some measure of statistical dependence as discussed above or in much more detail in Comon (1994) and subsequent literature. Since statistically independent components are necessarily uncorrelated, ICA usually proceeds by first computing the PCA, so that the spectral decomposition of $\widehat{\mathbf{K}}_{\boldsymbol{\alpha}}$ is known, and then applying a prewhitening transformation as in (8.67). At this point we have a set of sample vectors such that the sample covariance matrix is the unit matrix, and all further unitary transformations preserve this property. We thus decompose the matrix $\mathbf{M}$ as

$$\mathbf{M} = \mathbf{U}\widehat{\mathbf{K}}_{\boldsymbol{\alpha}}^{-\frac{1}{2}} , \tag{8.261}$$

where $\mathbf{U}$ is unitary. ICA amounts to choosing $\mathbf{U}$ so as to minimize the chosen measure of statistical dependence.

When ICA is carried out on training sets of natural scenes, the results are quite striking (see Bell and Sejnowski, 1997; Field, 1987; Olshausen and Field, 1996). The columns of $\mathbf{M}$ turn out to be localized, bandpass functions similar to wavelets or to the channels in the human visual system (a topic to be treated in more detail in Chap. 14), suggesting that humans may have evolved in such a way as to process natural scenes through statistically independent channels (see also Barlow, 1989).

One practical implication of the observation that the independent components are localized is that we can treat small pieces of the same object (or image) as independent samples. Bell and Sejnowski (1997), for example, consider 12×12 segments of an image as the samples on which they perform ICA. The resulting ICA filters are smaller than 12 pixels, even though the corresponding PCA filters tend to fill the 12×12 region. The authors note, however, that the restriction to such a small region may be an unrealistic feature of their approach. In addition, pixels themselves are unrealistic if we wish to draw conclusions about "natural scenes."

We shall revisit ICA in the context of texture analysis in Sec. 8.4.4. In that application, ICA is considerably simplified because textures are at least approximately stationary.

8.4.2 Multipoint densities

As we saw in Sec. 8.2.2, another kind of PDF for a random process is a collection of P-point densities of the form $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2), ..., f(\mathbf{r}_P)]$. In principle one needs densities like this for all P to completely characterize the process, but often we must be content with $P = 1$ and 2.

In a sense, multipoint densities are just special cases of the Hilbert-space densities discussed above. If we use delta functions as basis functions for the space (see Sec. 2.2.6), then $f(\mathbf{r}_p)$ is the coefficient α_p associated with basis function $\delta(\mathbf{r} - \mathbf{r}_p)$, and $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2), ..., f(\mathbf{r}_P)]$ is a P-dimensional marginal of a Hilbert-space density. This marginal is, however, a function of P spatial variables, so it is a richer description of the statistics of the random process than $\mathrm{pr}(\alpha_1, \alpha_2, ..., \alpha_P)$ would be with preselected basis functions.

If we have a means of computing $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2), ..., f(\mathbf{r}_P)]$, we can in principle do it for all values of each of the spatial arguments, but a less ambitious goal is to sample the function on a regular spatial grid, making it a discrete random process. If $\mathbf{r}$ is a qD vector and we sample each component to L values, then $\mathbf{f}$ is specified by $N = L^q$ numbers, and the full density is defined in an ND space. In this sampled case, therefore, all of the P-fold multipoint densities can be computed from the ND density on $\mathbf{f}$. Nevertheless, it may be computationally or conceptually simpler to compute the multipoint densities directly rather than as marginals of a high-dimensional multivariate density.

Pointwise evaluation of random functions Before analyzing multipoint densities in more detail, we have to deal with one mathematical subtlety. So far we have assumed only that each sample function $f(\mathbf{r})$ is in an $\mathbb{L}_2$ space, but we noted in Sec. 1.8 that not all functions in $\mathbb{L}_2$ are defined pointwise. If we want an expression like $f(\mathbf{r}_1)$ to be rigorously defined, we must assume that $f(\mathbf{r})$ lies in a reproducing-kernel Hilbert space (RKHS), which might be a subspace of $\mathbb{L}_2$. For imaging purposes, this restriction entails no loss of generality; we saw in Chap. 7 that the imaging operator $\mathcal{H}^\dagger\mathcal{H}$ is a nonnegative-definite Hermitian operator, and we know from Sec. 1.8.2 that such an operator can be used to define an RKHS. Assuming that $f(\mathbf{r})$ lies in that particular RKHS is equivalent to saying that we are discussing the statistics of the measurement component of the object, and that component is necessarily in an RKHS and hence defined pointwise.

The same conclusion can be reached by assuming that we are not interested in the statistics of an actual $f(\mathbf{r})$ but rather those of some linear approximation to it, such as the functions $f_a(\mathbf{r})$ or $f_t(\mathbf{r})$ discussed in Sec. 7.1.3. As we saw there, these functions lie in an RKHS called representation space, so they too can be defined pointwise. For example, we might construct a linear approximation by use of pixel functions, so $f_a(\mathbf{r}_1)$ would refer to the gray level[10] of a pixel centered at $\mathbf{r} = \mathbf{r}_1$.

In what follows we shall use the notation $f(\mathbf{r})$ but always with the implicit assumption that the function corresponds to a vector in an RKHS. Thus we might really mean $f_{meas}(\mathbf{r})$ or $f_a(\mathbf{r})$, but we shall omit subscripts for convenience. As a practical matter, essentially the only thing we rule out with this assumption is that $f(\mathbf{r})$ is white noise or some other generalized, infinite-energy random process.

Single-point PDFs For $P = 1$ and a fixed choice of $\mathbf{r}$, $\mathrm{pr}[f(\mathbf{r})]$ is a univariate PDF for the gray level $f(\mathbf{r})$ at point $\mathbf{r}$. This density can be represented as an ordinary 1D function as in Fig. 8.5. Of course, this function may depend in general on the choice of evaluation point $\mathbf{r}$, so it can also be plotted as a function of the Cartesian coordinates of $\mathbf{r}$, as shown in Fig. 8.6.

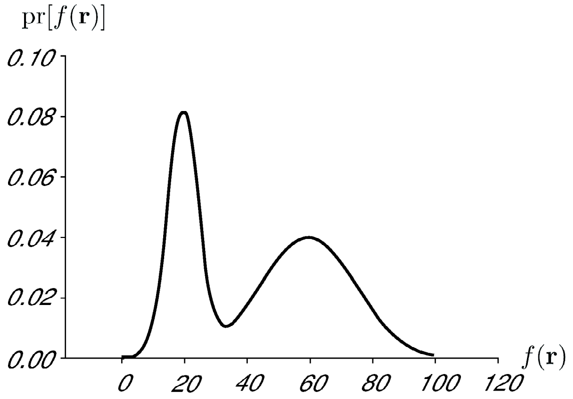


Fig. 8.5 Univariate PDF $\mathrm{pr}[f(\mathbf{r})]$ plotted as a function of $f(\mathbf{r})$ for fixed $\mathbf{r}$.

---

[10] Even though we are talking about pixels and gray levels here — terms often associated with images — we emphasize that we are discussing *object* models.

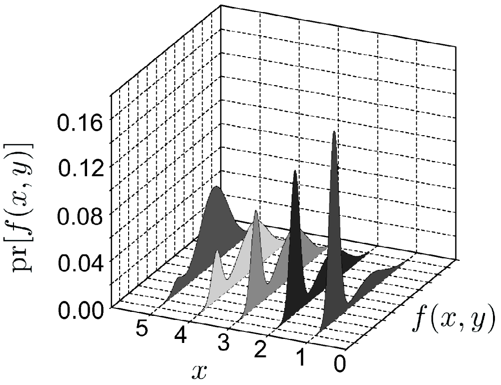


Fig. 8.6 Same PDF as in Fig. 8.5 but now plotted as a function of both $f(\mathbf{r})$ and $\mathbf{r}$.

Since it is univariate, $\mathrm{pr}[f(\mathbf{r})]$ admits of a frequentist interpretation and a histogram estimate. The considerations are essentially the same as for the univariate density $\mathrm{pr}(\alpha_n)$; if we have a source of object functions $f(\mathbf{r})$, such as a computer simulation code, we can evaluate each sample function at any chosen point, say $\mathbf{r} = \mathbf{r}_1$, and define a histogram estimate analogous to (8.256):

$$\widehat{\mathrm{pr}}_{f(\mathbf{r})}[f_k(\mathbf{r}_1)] = \frac{1}{\Delta f} \frac{J\left[f_k(\mathbf{r}_1) - \frac{1}{2}\Delta f \le f(\mathbf{r}_1) < f_k(\mathbf{r}_1) + \frac{1}{2}\Delta f\right]}{J}, \tag{8.262}$$

where the numerator is the number of sample objects for which the value $f(\mathbf{r}_1)$ falls in an interval of width Δf centered on the chosen value $f_k(\mathbf{r}_1)$, and J is the total number of samples. The number of bins in this histogram is just $f_{max}/\Delta f$, where f_{max} is the maximum value of $f(\mathbf{r})$. The k^{th} bin is centered on the point $f_k(\mathbf{r}_1)$ if

$$k = \frac{f_k(\mathbf{r}_1)}{\Delta f}. \tag{8.263}$$

For notational simplicity, we denote the numerator in (8.262) as J_k, which is just the observed number of samples in bin k, but we must keep in mind that the histogram is specific to the point $\mathbf{r}_1$.

The same statistical considerations apply here as in the last section. If the experiment is repeated many times with different sample objects, J_k will be approximately a Poisson random variable. The mean value of J_k will be J times the probability that the gray level will fall in bin k, or

$$\langle J_k \rangle = J \,\mathrm{Pr}\left[f_k(\mathbf{r}_1) - \tfrac{1}{2}\Delta f \le f(\mathbf{r}_1) < f_k(\mathbf{r}_1) + \tfrac{1}{2}\Delta f\right] \approx \mathrm{pr}_{f(\mathbf{r})}[f_k(\mathbf{r}_1)]\,\Delta f\,. \tag{8.264}$$

As in the previous section, we can construct a 100-bin histogram with a relative error of 10% in the value estimated for the k^{th} bin if $\langle J_k \rangle = 100$; a relative error of 1% requires $\langle J_k \rangle = 10^4$. If we assume that $\mathrm{pr}[f(\mathbf{r})]$ is relatively flat over the range from 0 to f_{max}, then we require $J = 10^4$ for 10% accuracy and 10^6 for 1% accuracy in a 100-bin. Again, these numbers are within the capabilities of modern computers if the sample objects are simulated.

One might think that we are far from characterizing the object random process even to order $P = 1$ since we have fixed the evaluation point at $\mathbf{r} = \mathbf{r}_1$ in the discussion above. In fact, however, once we have a source of sample objects $f(\mathbf{r})$, we can evaluate them at as many points as we please, and we can construct histogram estimates of $\mathrm{pr}[f(\mathbf{r})]$ on a grid of spatial points with very little increased effort. A 100×100 grid for a 2D object, for example, requires that we construct 10,000 histograms. If k ranges from 1 to 100 for each sample $\mathbf{r}$ and the observed value of J_k does not exceed 255, then we can store the results in just 1 Megabyte of memory.

As a semantic point, each of the histograms discussed above is a histogram of gray levels; it is not, however, what is usually called a gray-level histogram in the image-processing community. In that community, it is common to compute a histogram of the gray levels at *all points within a single image* for purposes of display manipulation or data compression. The histograms we are discussing here describe the distribution of gray levels *at a single point in an ensemble of images.* Where confusion may result, we shall distinguish between *single-image* histograms and *single-point* or ensemble histograms.

For stationary, ergodic random processes, the single-image histogram can be used in place of the ensemble histogram as an estimator of the single-point PDF, but these two histograms should not be equated in general. The single-image histogram can give a very biased estimate of the PDF if there is even a slight deviation from stationarity across the image. Consider, for example, the common situation where the mean gray level varies slowly across the image; in that case the single-image histogram can be much broader than the ensemble histogram at a fixed point and hence a fixed mean gray level.

Two-point PDFs For fixed $\mathbf{r}_1$ and $\mathbf{r}_2$, the two-point density $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]$ is a bivariate density on the two scalar random variables $f(\mathbf{r}_1)$ and $f(\mathbf{r}_2)$. This density can be represented by a 2D plot, where the axes are $f(\mathbf{r}_1)$ and $f(\mathbf{r}_2)$. A full characterization to order $P = 2$ requires evaluation of such bivariate densities for all $\mathbf{r}_1$ and $\mathbf{r}_2$ in $\mathbf{S}_f$.

The two-point density can also be interpreted in frequentist terms, though more sample objects are required than in the single-point case. If we again choose $f_{max}/\Delta f = 100$, then there are 10,000 bins in a histogram representing $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]$. A calculation similar to the one above shows that J must be about 10^6 for 10% accuracy and 10^8 for 1% accuracy if the underlying PDF is relatively flat. Moreover, 10^8 such histograms would be required if $\mathbf{r}$ is 2D and both $\mathbf{r}_1$ and $\mathbf{r}_2$ are sampled on 100×100 spatial grids, and 10 GB of storage would be needed to hold the results. In short, full experimental characterization of the random process to order P becomes rapidly more difficult as P increases.

The histogram approximation to the bivariate density $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]$ is related to, but not identical to, the *co-occurrence matrix* used in image processing and pattern recognition (Pratt, 1991). The distinction is the same as the one between single-point and single-image histograms. The co-occurrence matrix is a random matrix characteristic of a single image or a smaller region within a single image. It is a histogram of the joint occurrence of binned or quantized gray levels in that image. It is independent of absolute position within the region or image but it does depend on the relative position $\mathbf{r}_2 - \mathbf{r}_1$. The density $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]$, on the other hand, is a nonrandom characteristic of the ensemble and a function of two position vectors. A histogram approximation to $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]$ is also random since it is formed from a finite number of samples, but this randomness can in principle be reduced arbitrarily by letting the number of samples grow.

If each sample function is drawn from an ergodic random process (see Sec. 8.2.4), then the co-occurrence matrix computed from one sample function is also an estimator of $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]$.

Local models If $\mathbf{r}_1$ and $\mathbf{r}_2$ are far apart in an object, $f(\mathbf{r}_1)$ and $f(\mathbf{r}_2)$ might be statistically independent, or nearly so. For example, in a computed-tomography scan of the chest, the gray level at a point in the lungs would be expected to be independent of the gray level at a point in the spine. Two nearby points in the same lung would, however, be expected to be dependent. A stochastic model that takes account of this property is called a *local model.*

To see the structure of a local model, let us first consider two well-separated points. If the gray levels at these two points are statistically independent, the two-point PDFs are determined uniquely from the single-point ones:

$$\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)] = \mathrm{pr}[f(\mathbf{r}_1)]\,\mathrm{pr}[f(\mathbf{r}_2)]\,. \tag{8.265}$$

As discussed in Sec. C.1.6, the independence condition in (8.265) can also be written as

$$\mathrm{pr}[f(\mathbf{r}_1)|f(\mathbf{r}_2)] = \mathrm{pr}[f(\mathbf{r}_1)]\,. \tag{8.266}$$

Now consider a countable set of points, say on a regular lattice in object space. The gray level at some particular point $\mathbf{r}_k$ will often depend on the values at other points $\mathbf{r}_i$ provided they are close to the chosen point $\mathbf{r}_k$, but it could be statistically independent of the values at more distant points. We define the *neighborhood* $\mathcal{N}_k$ of the point $\mathbf{r}_k$ as the set of points close to $\mathbf{r}_k$ in this sense, and we denote the complete set of points in the object support as $\mathcal{S}$. Then a local statistical model is one for which [*cf.* (8.266)]

$$\mathrm{pr}[f(\mathbf{r}_k)|\{f(\mathbf{r}_i), \mathbf{r}_i \in \mathcal{S}, i \neq k\}] = \mathrm{pr}[f(\mathbf{r}_k)|\{f(\mathbf{r}_i), \mathbf{r}_i \in \mathcal{N}_k\}]\,, \tag{8.267}$$

where $\mathbf{r}_i \in \mathcal{N}_k$ is read "point $\mathbf{r}_i$ is an element of the set $\mathcal{N}_k$," or somewhat more colloquially, "$\mathbf{r}_i$ is a neighbor of $\mathbf{r}_k$." As we see from (8.267), the form of the marginal density on $f(\mathbf{r}_k)$ in a local model is determined fully by the values in the neighborhood $\mathcal{N}_k$, and points outside this neighborhood can be neglected for purposes of describing the statistics at $\mathbf{r}_k$.

A local model defined on a discrete lattice as in (8.267) is called a *Markov random field* or MRF. Developed by Besag (1973) and Cross and Jain (1983) for describing textures, MRFs have received considerable attention as Bayesian priors in image reconstruction (see Sec. 15.3.3), but relatively little effort has been expended on establishing their validity as empirical distributions in a frequentist sense. One exception is Herman and Chan (1995), who discussed so-called *image-modeling MRFs* where a sample drawn from the MRF density would have the same neighborhood statistics as the image (object) being modeled.

Regional models and mixture models Often objects can be divided into distinct regions with different statistical properties. In a chest radiograph, for example, the lungs are in more or less the same place for all patients, and the heart is generally situated below the left lung. Before seeing a particular patient's radiograph, we can define a region that is likely to contain lung and another that is likely to contain heart. Of course, this definition is not absolute; a collapsed lung or an enlarged heart, or simply normal variations in patient size and positioning, could mean that the *a priori* region assignment is incorrect. Various strategies are available for refining the region assignments, including image recentering and warping and various segmentation algorithms. None of these methods is perfect, however, and the best we can do is to assess the probability that a particular point is associated with a given region.

If we denote by $\mathcal{S}_i$ the set of points associated with region i, the univariate PDF on the gray level at point $\mathbf{r}$ is given by

$$\mathrm{pr}[f(\mathbf{r})] = \sum_i \mathrm{pr}[f(\mathbf{r})|\mathbf{r} \in \mathcal{S}_i] \, \mathrm{Pr}(\mathbf{r} \in \mathcal{S}_i) \, . \tag{8.268}$$

An analogous expression can be given for the two-point PDF:

$$\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)] = \sum_i \sum_k \mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)|\mathbf{r}_1 \in \mathcal{S}_i, \mathbf{r}_2 \in \mathcal{S}_k] \, \mathrm{Pr}(\mathbf{r}_1 \in \mathcal{S}_i, \mathbf{r}_2 \in \mathcal{S}_k) \, . \tag{8.269}$$

If gray levels in different regions are statistically independent, this equation becomes

$$\begin{aligned} &\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)] \\ &= \sum_i \sum_k (1 - \delta_{ik}) \, \mathrm{pr}[f(\mathbf{r}_1)|\mathbf{r}_1 \in \mathcal{S}_i] \, \mathrm{pr}[f(\mathbf{r}_2)|\mathbf{r}_2 \in \mathcal{S}_k] \, \mathrm{Pr}(\mathbf{r}_1 \in \mathcal{S}_i, \mathbf{r}_2 \in \mathcal{S}_k) \\ &\quad + \sum_i \mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)|\mathbf{r}_1 \in \mathcal{S}_i, \mathbf{r}_2 \in \mathcal{S}_i] \, \mathrm{Pr}(\mathbf{r}_1 \in \mathcal{S}_i, \mathbf{r}_2 \in \mathcal{S}_i) \, . \end{aligned} \tag{8.270}$$

Another special case is a *piecewise-constant model* where all points within a given region have the same gray level in each sample function of the random process, though that value (as well as the borders of the region) can vary randomly from one sample function to the next. In that case,

$$\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)|\mathbf{r}_1 \in \mathcal{S}_i, \mathbf{r}_2 \in \mathcal{S}_i] = \delta[f(\mathbf{r}_1) - f(\mathbf{r}_2)] \, \mathrm{pr}[f(\mathbf{r}_1)|\mathbf{r}_1 \in \mathcal{S}_i] \, . \tag{8.271}$$

The density in (8.268) is an example of a *mixture model* where the random quantity is divided into classes, and the overall density is a weighted sum of the densities for different classes. In (8.268), a class is identified with a spatial region, but other kinds of classes are important in imaging as well. In medical imaging, for example, different disease states are (we hope) described by different PDFs. Similarly, in aerial photography, crops, cities, oceans and forests would require different statistical models.

In such cases, the general form of the object PDF is

$$\mathrm{pr}(\mathbf{f}) = \sum_i \mathrm{pr}[\,\mathbf{f}|\text{class } i\,] \, \mathrm{Pr}(\text{class } i) \, . \tag{8.272}$$

The key difference between (8.268) and (8.272) is that the former applies to a univariate density at a specific point $\mathbf{r}$, while the latter is a general statement applying to the entire density of the process.

Specifically, if we represent $\mathbf{f}$ by an $N \times 1$ coefficient vector $\boldsymbol{\alpha}$, the mixture density (8.272) takes the form

$$\mathrm{pr}(\boldsymbol{\alpha}) = \sum_i \mathrm{pr}(\boldsymbol{\alpha}|\text{class } i) \, \mathrm{Pr}(\text{class } i) \, , \tag{8.273}$$

and the marginal on a single component of $\boldsymbol{\alpha}$ is

$$\mathrm{pr}(\alpha_n) = \sum_i \mathrm{pr}(\alpha_n|\text{class } i)\,\mathrm{Pr}(\text{class } i)\,. \tag{8.274}$$

8.4.3 Normal models

The basic properties of normal random processes and random vectors were introduced in Sec. 8.3. Here we revisit normal models with the goal of understanding when and how they apply specifically to the statistical description of objects.

When it is possible to use normal models in imaging, a considerable mathematical simplification results. As we saw in Sec. 8.3, the PDF for a normal random vector is fully determined by the mean vector and the covariance matrix. Moreover, any linear transformation of a normal random vector leaves it normal, so a full analysis of the effect of a linear operator requires only that we transform the mean and covariance, using simple formulas developed in Sec. 8.1.5.

These properties of normal random vectors extend readily to normal random processes. The full PDF of any random process is infinite-dimensional, but in the normal case we can take advantage of the fact that any marginal or conditional density derived from a normal PDF, even an infinite-dimensional one, is also normal. Thus if we choose to describe a normal random process by Hilbert-space marginal densities of the form $\mathrm{pr}(\alpha_1, \alpha_2, ..., \alpha_P)$ or by multipoint densities like $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2), ..., f(\mathbf{r}_P)]$, we can be assured that these densities will all be normal.

Central limits To establish the validity of a normal model, we must usually argue that the central-limit theorem applies, as it does when independent random variables or vectors are added together. One way this can happen is when a pixel or voxel representation is used for the object, and subregions of the pixel or voxel are statistically independent. As an example, consider an airborne optical camera viewing a meadow. The camera does not resolve individual blades of grass, and an adequate 2D object representation can use a pixel that covers many blades. It is reasonable to argue that the blades reflect light independently, so the total reflected light in one pixel tends to a normal distribution, at least when we consider only meadows and do not include, say, forests or beaches.

A somewhat more subtle example is nuclear medicine imaging of perfusion patterns in the lungs. In this technique, radioactive albumin particles are injected into a vein and get trapped in the alveoli (the functional units of the lungs where blood becomes oxygenated). The distribution of the trapped tracer is indicative of the perfusion of the lung, and it is this distribution that we regard as the object. Since nuclear medicine systems have very poor spatial resolution compared to the size of alveoli, we can choose a voxel size that contains many alveoli, and the voxel value is the sum of the activities in many alveoli. It is reasonable to presume that these activities are statistically independent, at least when one particular patient is considered. If we were to consider an ensemble of patients, some would have higher perfusion in a particular region than others, and all alveoli in this region would tend to fluctuate together; we avoid this kind of dependence by conditioning the PDF on a particular patient and hence a particular perfusion pattern. In a frequentist sense, this conditional PDF describes the hypothetical distribution that would result from making many different injections of albumin particles into a single patient.

Gaussian mixture models In the two examples just given to justify use of the central-limit theorem, we had to be careful to restrict the ensemble of objects under consideration. In the aerial photography example, we had to consider only meadows and not forests or beaches, and in the nuclear medicine example we had to consider repeated injections into one patient rather than a more realistic ensemble of patients.

To analyze a broader ensemble, we do not necessarily have to abandon the central-limit theorem; instead, we can divide the different objects (or different regions of the same object) into classes and use a mixture density as in (8.272). If we can argue that a normal PDF applies to each component of the mixture, then the resulting model is called a *Gaussian mixture model.*

If $\boldsymbol{\alpha}$ is conditionally multivariate normal for each class, then α_n is conditionally univariate normal, so $\mathrm{pr}(\alpha_n|\text{class } i)$ in (8.274) is fully specified by the conditional mean $\overline{\alpha}_{ni}$ and the conditional variance σ_{ni}^2:

$$\mathrm{pr}(\alpha_n) = \sum_i \frac{1}{\sqrt{2\pi\sigma_{ni}^2}} \exp\left[-\frac{(\alpha_n - \overline{\alpha}_{ni})^2}{2\sigma_{ni}^2}\right] \mathrm{Pr}(\text{class } i)\,. \tag{8.275}$$

If we must use a large number of classes in order to justify the normal law for each class, it might be better to consider a continuum of classes and write

$$\mathrm{pr}(\alpha_n) = \int_{-\infty}^{\infty} d\overline{\alpha}_n \int_0^{\infty} d\sigma_n^2 \; \mathrm{pr}(\overline{\alpha}_n, \sigma_n^2) \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{(\alpha_n - \overline{\alpha}_n)^2}{2\sigma_n^2}\right]\,. \tag{8.276}$$

Similarly, the multivariate density on $\boldsymbol{\alpha}$ for a discrete set of classes is

$$\mathrm{pr}(\boldsymbol{\alpha}) = \sum_i \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{K}_i)}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}}_i)^t \mathbf{K}_i^{-1} (\boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}}_i)\right] \mathrm{Pr}(\text{class } i)\,, \tag{8.277}$$

where $\overline{\boldsymbol{\alpha}}_i$ and $\mathbf{K}_i$ are, respectively, the mean vector and covariance matrix for $\boldsymbol{\alpha}$ under class i. For a continuum of classes, we can write

$$\mathrm{pr}(\boldsymbol{\alpha}) = \int_{\infty} d^N\overline{\alpha} \int_{\infty} d\mathbf{K} \; \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{K}_i)}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}})^t \mathbf{K}^{-1} (\boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}})\right]\,, \tag{8.278}$$

where $d\mathbf{K}$ is a shorthand for the differential of all components of $\mathbf{K}$.

No matter which of these mixture formulas we use, we do not expect the resulting PDF to be normal. For example, in the simple case of the univariate expression (8.275) with just two classes, we would get a bimodal PDF with one Gaussian peak for each class.

High-pass and band-pass filters There are many circumstances where we either pass an image through a high-pass or band-pass spatial filter or consider an object to consist of a superposition of such components. For example, edges in an image are often detected with some sort of derivative filter, and derivatives suppress the DC component[11] of the image. Other examples of filters with zero DC response include

---

[11] The common jargon, *DC component*, does not, of course refer to direct current. Instead it implies zero spatial frequency, by analogy to the zero temporal frequency of a steady current. In coherent optical processing, the Fourier transform of an object is displayed as an optical amplitude distribution centered on the optical axis of a lens system, and in that case it has been suggested that *DC* stands for *dot in the center.*

wavelets (see Sec. 5.3), channels in the human visual system (see Sec. 14.2) and filters used to extract discrete cosine transforms (except, of course, the DC term in the transform). Continuous objects can be represented by zero-DC components, for example in the Fourier-series basis of (7.13), a wavelet basis or a basis of Gabor functions (see Sec. 5.1.4). As we noted in Sec. 8.4.1, approximately independent components can be obtained by filtering with localized band-pass filters.

In all of these cases, an expansion coefficient is computed by forming a scalar product of the object function with a zero-DC function. For both objects and images, therefore, it is of considerable interest to have a stochastic model for the output of a high-pass filter.

In Sec. 8.3.3 we showed that linear filtering of a Gaussian random process yields a Gaussian random process, so if the input to a filter is Gaussian, the output must be also. It has been observed empirically, however, many images have a decidedly non-Gaussian distribution results after high-pass or band-pass filtering. As seen in the example in Fig. 8.7, the gray-level histograms are typically sharply peaked around zero and display long tails (Heine *et al.*, 1999; Bell and Sejnowski, 1997). In statistical lingo, these histograms have a large kurtosis. As defined in (C.41), the kurtosis for a Gaussian is 3 (though many books subtract off the 3 and make the kurtosis of a Gaussian 0), and gray-level histograms following high-pass filtering often have kurtosis substantially larger than 3. Statistical pedants refer to such distributions as *leptokurtic* (Greek *lepto*, thin or fine); the opposite condition, kurtosis less than that of a Gaussian, is referred to as *platykurtic* (Greek *platys*, broad or flat — behold the platypus!).

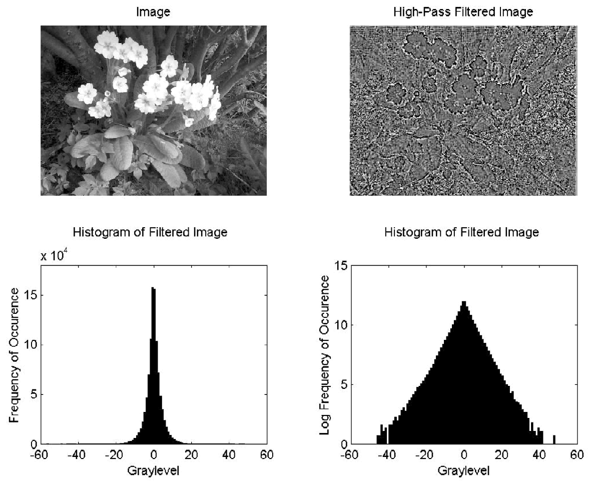


Fig. 8.7 *Top*: A typical image before and after high-pass filtering. *Bottom*: Gray-level histogram of the high-pass filtered image (note that the right plot is vs. log frequency of occurrence).

Filtering of Gaussian mixtures Heine *et al.* (1999) offered an explanation for high kurtosis after wavelet filtering, but it made assumptions about scale-invariance that were specific to wavelets. Lam and Goodman (2000) derived the PDF of the coefficients in a discrete cosine transform from a Gaussian mixture model. Clarkson and Barrett (2001) extended that argument and showed that kurtotic distributions were an inevitable consequence of high-pass or band-pass filtering of Gaussian mixtures; we shall sketch here the derivation given by Clarkson and Barrett.

If we think of high-pass filtering as a convolution, then the output is a scalar product of the shifted kernel function with the input. If the kernel contains both positive and negative components, we can suppress the shift variable and write the output for one position of the kernel as

$$z = u - v\,, \tag{8.279}$$

where u arises from the positive part of the filter and v from the negative part. This equation applies whether we think of the input to the filter as a random process or a random vector in a pixel representation. Moreover, it applies also to computation of an expansion coefficient in a representation where the expansion function has positive and negative parts.

We expect u and v to be highly correlated since they come from the same region of the input, but it is reasonable to assume that they have the same mean if the filter has zero DC response. This conclusion follows rigorously if we can assume that all points within the region spanned by the kernel (at a specific shift) have the same mean, and it may also be a good approximation even with a space-variant mean since it requires only that the spatial average of the mean over the positive regions of the kernel equal that over the negative regions. (Consider a difference-of-Gaussians filter, where a positive central peak is surrounded by a negative ring; the means of u and v will be equal if the spatial average of the input mean in the negative ring is the same as the spatial average in the central peak.) Thus we assume

$$\overline{u} = \overline{v}\,; \qquad \overline{z} = 0\,. \tag{8.280}$$

Note that the overbar here implies an ensemble mean; it has nothing to do with spatial averages. We make no assumptions about stationarity or ergodicity, and there is no implication that ensemble averages can be approximated by spatial ones.

Now let us assume that u and v are drawn from a Gaussian mixture. To see the essential results, we assume first that u and v are *conditionally* uncorrelated, for any one component of the mixture, so that the entire correlation between the two variables results from averaging over components in the mixture. Similarly, we assume that u and v have the same conditional variances, so in fact they are conditionally i.i.d. These assumptions may not always be justified, and they will be relaxed below; for now, we write

$$\mathrm{pr}(u, v|\sigma, m) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(u-m)^2 + (v-m)^2}{2\sigma^2}\right]\,, \tag{8.281}$$

where m is the common mean of u and v, and σ is the common standard deviation.

The corresponding conditional density on z is given by

$$\mathrm{pr}(z|\sigma, m) = \int_{-\infty}^{\infty} du\; \mathrm{pr}(u, u - z|\sigma, m)\,. \tag{8.282}$$

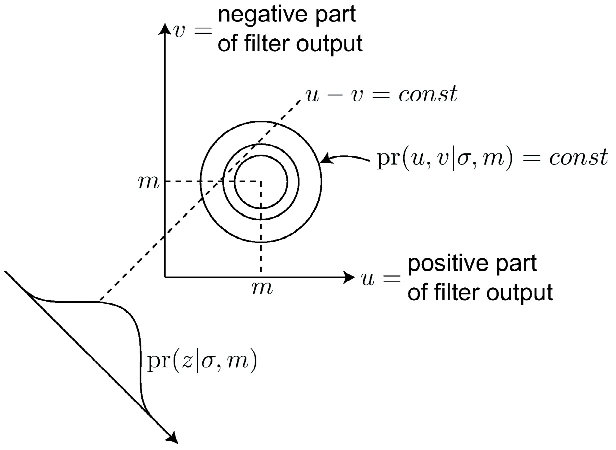

$v =$ negative part of filter output
$u - v = const$
$\mathrm{pr}(u, v|\sigma, m) = const$
m
m
$u =$ positive part of filter output
$\mathrm{pr}(z|\sigma, m)$

Fig. 8.8 Illustration of the integral encountered in analyzing Gaussian mixture models.

As illustrated in Fig. 8.8, this integral can be interpreted as a 1D projection or Radon transform (see Sec. 4.4) of the 2D function $\mathrm{pr}(u, v|\sigma, m)$. We see graphically that the result is independent of m, and by completing the square we obtain

$$\mathrm{pr}(z|\sigma) = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} du \, \exp\left[-\frac{u^2 + (u-z)^2}{2\sigma^2} \right] = \frac{1}{2\sigma\sqrt{\pi}} \exp\left(-\frac{z^2}{4\sigma^2} \right) . \quad (8.283)$$

Note that we have written this density as conditional on the standard deviation σ rather than the variance σ^2. We are free to choose either, but the standard deviation is convenient when we write out the overall density on z. Since the conditional mean does not influence the statistics of z, the mixture can be specified by a univariate prior on σ, and we find

$$\mathrm{pr}(z) = \frac{1}{2\sqrt{\pi}} \int_0^{\infty} \frac{d\sigma}{\sigma} \exp\left(-\frac{z^2}{4\sigma^2} \right) \mathrm{pr}(\sigma) . \quad (8.284)$$

By comparison with (4.85), we recognize (8.284) as a Mellin convolution, and many interesting properties of $\mathrm{pr}(z)$ follow from this observation. Since Mellin transforms convert Mellin convolutions into products, and since Mellin transforms can be interpreted as moments (see Sec. 4.2.2), it follows that moments of z are related simply to moments of σ; from Clarkson and Barrett (2001), the relation is

$$\langle z^k \rangle = \frac{2^k \Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi}} \langle \sigma^k \rangle , \quad (8.285)$$

where $\Gamma(\cdot)$ is the gamma function.

From this moment relation and a little algebra, we find

$$\langle z^4 \rangle - 3\langle z^2 \rangle^2 = 12[\langle \sigma^4 \rangle - \langle \sigma^2 \rangle^2] . \quad (8.286)$$

The kurtosis, defined as $\langle z^4 \rangle / \langle z^2 \rangle^2$, is 3 for a Gaussian, so the left-hand side of this expression would be zero for a Gaussian. By the Schwarz inequality, however, the right-hand side is ≥ 0, so z always has a kurtosis greater than or equal to that of a Gaussian, with equality if and only if $\mathrm{pr}(\sigma)$ is a delta function. In short, leptokurtic distributions are inevitable when a Gaussian mixture is filtered with a high-pass or band-pass filter. Moreover, the resulting densities for z often take simple, symmetric forms, quite robust to the detailed assumptions about $\mathrm{pr}(\sigma)$.

Several different analytical forms have been suggested as empirical descriptions of long-tailed densities like those shown in Fig. 8.8. When there is a sharp cusp at the origin, a natural choice is the Laplace or double-exponential density. A family of densities intermediate between Laplace and Gaussian can also be defined with $\mathrm{pr}(z) \propto \exp(-a|z|^p)$, so $p = 1$ is the Laplace density and $p = 2$ is the Gaussian. The parameters p and a can be adjusted to fit empirical densities. Another option is the Lévy family, defined not by the density but by the characteristic function, which has the form $\psi(\xi) = \exp(-b|\xi|^q)$. The corresponding densities cannot be stated as simple analytic functions except when $q = 2$, which is the Gaussian, and $q = 1$, which is the Cauchy density (see Sec. C.5.10). Again, q and b can be treated as adjustable parameters.

Mixtures of correlated Gaussians So far we have considered only a specific Gaussian mixture where u and v were i.i.d. normal, but the result can readily be generalized. Suppose u and v are bivariate normal with a covariance matrix of the form

$$\mathbf{K}_{uv} = \begin{bmatrix} a & b \\ b & c \end{bmatrix} . \tag{8.287}$$

As the reader may show, (8.284) is still valid with this model, only now σ^2 is not a univariate variance but rather $\frac{1}{2}(a + c - 2b)$ (see Clarkson and Barrett, 2001). Thus the initial assumption that u and v are i.i.d. has no essential effect on the conclusions.

Normals and entropy It is not always necessary to invoke the central-limit theorem in order to arrive at a normal probability law. It occurs also in a Bayesian context when one has partial information about a distribution and wishes to complete the description as noncommittally as possible. One way to do this is to use the principle of maximum entropy. A critique of this approach in the context of image reconstruction is given in Sec. 15.3.3, but here we can be content to paraphrase Zhu *et al.* (1998): Entropy is a measure of randomness, and we should choose the density that is as random as possible in all unobserved dimensions and does not attempt to represent information that we do not have.

If we know the mean and variance of a random variable (or mean vector and covariance matrix of a random vector), these moments serve as constraints on the density, and we would like to find the density of maximum entropy consistent with these constraints. We shall carry through the calculation in the univariate case and simply state the multivariate result.

Consider a random variable x and suppose we know that its mean is $\overline{x}$ and its variance is σ^2. According to the principle of maximum entropy, we must choose $\mathrm{pr}(x)$ to maximize $\int_{-\infty}^{\infty} dx\ \mathrm{pr}(x) \ln \mathrm{pr}(x)$, subject to the constraints

$$\int_{-\infty}^{\infty} dx\ \mathrm{pr}(x) = 1 \, ; \qquad \int_{-\infty}^{\infty} dx\ x\ \mathrm{pr}(x) = \overline{x} \, ; \qquad \int_{-\infty}^{\infty} dx\ x^2\, \mathrm{pr}(x) = \sigma^2 + \overline{x}^2 \, . \tag{8.288}$$

The maximization can be performed by the method of Lagrange multipliers. We require that the Lagrangian functional,

$$L\{\mathrm{pr}(x)\} \equiv \int_{-\infty}^{\infty} dx\ \mathrm{pr}(x) \ln \mathrm{pr}(x) + \alpha \left[\int_{-\infty}^{\infty} dx\ \mathrm{pr}(x) - 1 \right]$$

$$+ \beta \left[\int_{-\infty}^{\infty} dx\ x\ \mathrm{pr}(x) - \overline{x} \right] + \gamma \left[\int_{-\infty}^{\infty} dx\ x^2\, \mathrm{pr}(x) - (\sigma^2 + \overline{x}^2) \right] , \qquad (8.289)$$

be unchanged by small perturbations of $\mathrm{pr}(x)$. Here, α, β and γ are the Lagrange multipliers, to be fixed by the constraint equations. If we perturb $\mathrm{pr}(x)$ by a small amount $\eta(x)$, and retain only terms linear in the perturbation, we find

$$L\{\mathrm{pr}(x) + \eta(x)\} - L\{\mathrm{pr}(x)\} = \int_{-\infty}^{\infty} dx\ \eta(x) \left\{ 1 + \ln \mathrm{pr}(x) + \alpha + \beta x + \gamma x^2 \right\} = 0 . \qquad (8.290)$$

Since $\eta(x)$ is arbitrary, this equation can hold only if the quantity in braces in the integrand is zero, so $\mathrm{pr}(x)$ must take the form

$$\mathrm{pr}(x) = \exp\left(-1 - \alpha - \beta x - \gamma x^2 \right) . \qquad (8.291)$$

Both this form and the constraints are satisfied if

$$\mathrm{pr}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \overline{x})^2}{2\sigma^2} \right] . \qquad (8.292)$$

Thus, if all we know about a random variable is its mean and variance, the maximum-entropy choice for its density is a normal. A similar calculation shows that if all we know about a random vector is its mean vector and covariance matrix, the maximum-entropy density is multivariate normal.

Positivity Appealing though normal distributions may be, they have one serious deficiency in many imaging applications. If the random variable or vector in question is inherently nonnegative, as physical objects often are, then the normal law cannot be strictly correct; it always predicts some finite probability of negative values. We shall now discuss several possible fixes for this problem.

One simple fix is just to consider situations where the standard deviation of the random variable is small compared to its mean; then the probability of getting a negative value is small and can perhaps be neglected without serious error. For the normal law to represent a nonnegative object, in particular, we must consider low-contrast scenes where the variation we are trying to describe is small compared to some spatial-average value. Such situations arise often in medical imaging or other applications involving a faint object on a bright background. They can be particularly useful for local statistical descriptions where the background may vary substantially over the whole scene but relatively little over a region of interest.

Another approach is to use a truncated Gaussian which is not allowed to go negative. Perhaps surprisingly, this is the maximum-entropy choice if we know the mean and variance of a random variable and also know that it is nonnegative. Retracing the calculation above, we see that (8.290) still holds with the simple modification of setting the lower limit of integration to zero, and (8.291) holds without modification. A more substantial modification does occur in (8.292), which can now be written as

$$\mathrm{pr}(x) = N \exp\left[-\frac{(x-x_0)^2}{2v^2}\right] \mathrm{step}(x)\,, \tag{8.293}$$

where N is a normalizing constant, x_0 is *not* the mean and v^2 is *not* the variance. Instead these quantities must be determined by numerically solving constraint equations like (8.288) but with a lower integration limit of 0.

Similarly, if we know the mean and covariance of a nonnegative random vector, a truncated multivariate normal is the maximum-entropy density. Again, however, the known mean and covariance cannot simply be plugged into the standard multivariate normal form.

Log-normals Another solution to the positivity problem is to use log-normals rather than normals. Since a log-normal is a density for a random variable whose log is normal, it is defined for any nonnegative variable, and the density is taken to be zero for negative values of the variable.

The density for a univariate log-normal is given in Sec. C.5.9 of App. C; the corresponding multivariate form is

$$\mathrm{pr}(\mathbf{f}) = \left[\prod_i \frac{1}{\sqrt{2\pi} f_i}\right] \frac{1}{\sqrt{\det(\mathbf{K})}} \exp\left\{-\tfrac{1}{2}\left[\ln(\mathbf{f}) - \boldsymbol{\mu}\right]^t \mathbf{K}^{-1} \left[\ln(\mathbf{f}) - \boldsymbol{\mu}\right]\right\}\,, \tag{8.294}$$

where the logarithm is to be interpreted componentwise, and $\boldsymbol{\mu}$ and $\mathbf{K}$ are the mean vector and covariance matrix of the Gaussian random vector $\ln \mathbf{f}$, not $\mathbf{f}$ itself. The reader may test her understanding of transformations of random variables by showing that the log of $\mathbf{f}$ is indeed a multivariate normal.

Often we can argue on physical grounds that the PDF for an object or image should tend to a log-normal. Consider, for example, transmission x-ray imaging of a thick, inhomogeneous 3D object. The 3D object can be divided into slabs, and the overall transmission of the object is the product of the slab transmissions. If the transmission of each slab is a random process, then the log of the product of the slab transmissions is also a random process, and if the individual slabs are statistically independent, then the log of the product is the sum of logarithms of independent random processes. Thus, regardless of the statistics of the slab log-transmissions, the overall log-transmission tends to a normal by the central-limit theorem, and hence the overall transmission itself tends to a log-normal.

In other situations as well, we can decompose the object function into a product of independent random variables. For example, in nuclear medicine we might inject a radioactive tracer into the blood stream and watch its migration through the circulatory system to some target organ. At each branching of the blood vessels, a tracer molecule can go in one of two directions, and if we consider a point in the vasculature after many branchings, then the number of molecules arriving there is the injected number times a product of a large number of random variables, one for each branch. The central-limit theorem suggests that the log of the tracer concentration at this point is normally distributed, so the concentration itself is log-normal.

There is an essential difference between log-normals and truncated normals as densities for nonnegative random variables or vectors. As the examples above suggest, we can expect the log-normal to be experimentally verifiable, in principle, so it can be interpreted in a frequentist sense. If the variable of interest is a product of many independent random variables, and each experiment results in different values for the individual variables, we can repeat the experiment many times, and the resulting histogram estimate of the density for the product variable will tend to a log-normal, and indeed this distribution is frequently observed experimentally.[12]

There is, in fact, a frequentist rationale for maximum entropy, and it will be sketched in Sec. 15.3.3, but it conceives of the object being constructed by throwing imaginary grains or blobs of gray level; it definitely does not suggest a concrete physical experiment. Thus, even though the truncated normal might be a maximum-entropy density, we should not expect to encounter it as the limit of an experimental histogram. Maximum entropy, as we used it above, is merely a way of going from known moments to a noncommittal PDF.

8.4.4 Texture models

A significant portion of the image-science literature deals with analysis, synthesis, recognition and segmentation of *textures*, defined loosely as spatial random fields with some degree of stationarity. Sometimes the stationarity is periodic, with basic repeating elements such as bricks in a wall or fibers in a woven[13] fabric. Sometimes it is continuous, as with a stucco wall rather than a brick one or the surface of the ocean, where the light reflected from the object can be described as a stationary random process. Sometimes the stationarity is only approximate, in one of the senses discussed in Sec. 8.2.4; the correlation properties might vary slowly, or they might be stationary only within some region boundaries. Sometimes, in fact, the stationarity is purely visual; two regions are said to be the same texture simply because a human observer cannot tell them apart.

Since textures are essentially stationary random processes, Fourier analysis is an important tool for analyzing them. We shall therefore start this section with a discussion of the role of Fourier analysis and power spectral densities, and then we shall briefly discuss methods for estimating power spectra.

Even when stationarity is a good approximation, an autocorrelation function or power spectral density may not capture all of the essential properties of a texture field. It may be necessary to specify also some aspects of the multivariate PDF in order to adequately describe a texture, and we shall describe several means of doing so.

Throughout this section we shall discuss not only methods of characterizing texture as a random process, but also methods for generating sample functions of the random process. An excellent general reference on methods of constructing sample functions with specified correlation properties and marginal distributions is Johnson (1994).

---

[12] Above we presented an argument that the total amount of tracer in a voxel should tend to a normal, and here we argue that the concentration at a point should be log-normal. These two arguments are not necessarily inconsistent, since a sum of log-normals can converge to a normal, but in fact this convergence is very slow (Barakat, 1976). Which distribution is actually observed is best resolved empirically.

[13] Texture comes from the Latin *texere*, to weave, so a fabric is the prototype of a texture.

Fourier Phase and magnitude Any spatial pattern, whether regarded as a deterministic function or as a sample function of a random process, is completely specified by its Fourier transform. This (continuous or discrete) Fourier transform is complex, but the modulus and phase convey essentially different information about the object. Fourier phase tells you where things are — if the position of an object is shifted, the phase changes but the modulus does not. Fourier modulus, on the other hand, tells you only how strongly different spatial frequencies contribute to the object.

In many cases, Fourier phase is more important than Fourier modulus in conveying the essence of an object. In a famous experiment, Oppenheim and Lim (1981) Fourier-transformed two images, one of the television news anchor Walter Cronkite and one of a clock. They then interchanged the Fourier phases, putting Walter's phase with the clock modulus and vice versa. After inverse transformation, the image with Walter's Fourier phase still looked like Walter, and the one with the phase of a clock looked like a clock.

With textures, on the other hand, the situation can be reversed. In a stationary random process we do not care where things are. One location is as good as another, at least statistically, so Fourier phase is much less important than Fourier modulus. Two stationary random processes with the same modulus but different phases are recognized as sample functions of the same texture. One common way of synthesizing sample textures, therefore, is to generate samples of white noise and pass them through a linear filter.

As an example, Bochud *et al.* (1999b) examined the relative importance of Fourier amplitude and phase in describing coronary angiograms (x-ray images of blood vessels after injection of an x-ray-absorbing material into the blood stream). In agreement with the remarks above, they found that the phase was important for describing the vessel, but not for the random anatomical background against which the vessel was seen. Though the background was not rigorously stationary, they showed that realistic images could be simulated by filtering white noise through a space-variant filter.

Estimation of power spectra or autocorrelation functions of images Suppose we have one or more sample images, and we want to generate additional images with similar texture by filtering white noise. To the extent that the texture is a stationary random process, we need to know the power spectral density or the stationary autocorrelation function. There is a large literature on estimating these quantities from sample images, and we confine ourselves here to a few general observations.

In Sec. 8.2.5 we mentioned — and dismissed — an apparently obvious approach to spectral estimation, the periodogram of a single sample image. Figure 8.1 illustrates the difficulty with this approach. Mallat (1999) refers to periodogram analysis as "naive spectral estimation;" one meaning of *naive* is "lacking information, uninformed" and the periodogram is naive in the sense that it does not incorporate prior information or beliefs into the spectral estimate. We certainly do not believe that the rapid fluctuations seen in Fig. 8.1 are meaningful features of the power spectrum (or if we did, we would need only to repeat the experiment to change our belief system). The situation is very similar to image reconstruction, discussed in much more detail in Chap. 15, where naive attempts at inverse filtering yield large fluctuations in the reconstructed image.

The Bayesian approach to this problem, in both image reconstruction and spectral estimation, is to define a prior probability on the function being estimated and then to seek an estimate consistent with both the data and this prior. In the Bayesian community, a preferred prior is the entropy, and maximum-entropy reconstructions do indeed eliminate the rapid fluctuations and yield smooth estimates. The details of this procedure, in the context of image reconstruction, are given in Sec. 15.3.3.

As we shall also see in Chap. 15, there are many other approaches, referred to collectively as *regularization*, that can be used to suppress fluctuations in reconstructed images, and each of these methods has its analog in spectral estimation. Many of these methods can also be described as Bayesian, but with priors other than entropy (see Sec. 15.3.3); all of them attempt to enforce our prior belief that the function being reconstructed (power spectrum or image) is smooth in some sense.

One way to enforce smoothness in spectral estimation is to model the spectrum as a smooth function with unknown parameters and then to estimate the parameters. For example, we could model the spectrum as a Gaussian and estimate its width, or as a Gaussian times a polynomial and estimate the polynomial coefficients also. One popular model, especially for time-series analysis, is the *autoregressive, moving average* or ARMA model where the spectrum is modeled as a ratio of polynomials (Oppenheim and Schafer, 1989).

Another model is to assume that the power spectrum varies as a power law, $\rho^{-\beta}$, and then to estimate the exponent β. Many images exhibit this behavior in practice (even when there is no reason to assume stationarity), and β is a useful phenomenological descriptor. Physical mechanisms that lead to power-law power spectra in the context of electrical noise are surveyed in Sec. 12.2.3.

Which regularization method is chosen depends on what one wants to do with the spectral estimate. If we want to simulate images that appear realistic to a human observer, we can use one of the psychophysical tests detailed in Sec. 14.2.3 to measure how well the observer can distinguish real texture images from white noise filtered with the estimated spectrum. If the real and simulated images are indistinguishable, it means that the estimated spectrum is good enough for this purpose; on the other hand, if they are readily distinguishable, it may mean that the spectral estimate is poor, or it may mean that the texture is more complicated than just filtered noise.

For many purposes, however, we need more than just visual realism. In texture recognition or discrimination, for example, we need a stochastic model in order to design an optimal discriminant function (see Sec. 13.2.12). If we use a Gaussian model, we need to know the inverse of the covariance matrix, and if we also assume stationarity, that means we need to know the reciprocal of the power spectrum. Even if the spectral estimate accurately represents the actual spectrum for the spatial frequencies where the spectrum is large, it may be a poor estimate in the tails and hence a poor estimate of the reciprocal spectrum. The best spectral estimate in this case is the one that leads to the best discrimination performance for a discriminant function based on the estimated spectrum (but tested on real images — not ones simulated from the estimated spectrum!).

As another example, we shall see in Chaps. 13 and 14 that some important measures of image quality are expressed in terms of the image power spectrum. If we do not know the actual spectrum, we must estimate it, and the adequacy of the spectral estimate must be judged by the accuracy of the corresponding estimates of figures of merit for image quality.

Estimation of power spectra or autocorrelation functions of objects Above we stated our goal as estimation of the power spectral density or autocorrelation function of a set of images. Often, however, what we really want to know is the power spectral density or autocorrelation function of the objects that formed the images.

Suppose we have a set of sample images $\{\mathbf{g}_j\,, j = 1, ..., J\}$, where the j^{th} image is related to an object $\mathbf{f}_j$ by $\mathbf{g}_j = \boldsymbol{\mathcal{H}}\mathbf{f}_j + \mathbf{n}_j$. We must assume that $\mathbf{f}_j$ is a sample function of a stationary (or at least quasistationary) random process in order to define an object power spectral density $S_f(\boldsymbol{\rho})$, and we need knowledge of $\boldsymbol{\mathcal{H}}$ and of the noise statistics in order to estimate $S_f(\boldsymbol{\rho})$.

As a simple example, suppose the imaging system is well approximated as a convolution (a CC LSIV system in the language of Sec. 7.2.3). Then $\mathbf{g}_j$ is a sample function of a stationary random process, and its power spectrum is denoted by $S_g(\boldsymbol{\rho})$. If we also assume that $\mathbf{n}_j$ is a sample function of a stationary random process, with power spectrum $S_n(\boldsymbol{\rho})$, then use of (8.156) shows that the image power spectrum is given by

$$S_g(\boldsymbol{\rho}) = |H(\boldsymbol{\rho})|^2\, S_f(\boldsymbol{\rho}) + S_n(\boldsymbol{\rho})\,. \tag{8.295}$$

The image spectrum $S_g(\boldsymbol{\rho})$ can be estimated by any of the methods suggested above, and the result can be denoted as $\widehat{S}_g(\boldsymbol{\rho})$. If we know the noise spectrum $S_n(\boldsymbol{\rho})$ independently from the physics of the imaging problem, then one reasonable estimate of the object spectrum is

$$\widehat{S}_f(\boldsymbol{\rho}) = \frac{\widehat{S}_g(\boldsymbol{\rho}) - S_n(\boldsymbol{\rho})}{|H(\boldsymbol{\rho})|^2}\,. \tag{8.296}$$

This method gives little information about $S_f(\boldsymbol{\rho})$ at frequencies for which $H(\boldsymbol{\rho})$ is small, and large errors in $S_f(\boldsymbol{\rho})$ can result from small errors in either $S_n(\boldsymbol{\rho})$ or $\widehat{S}_g(\boldsymbol{\rho})$. Moreover, the whole approach depends on modeling the system as CC LSIV and the noise as stationary.

A better approach is to use some parametric description of the object power spectrum, perhaps one that allows quasistationarity, and then to estimate the parameters from the data. This way, the system operator $\boldsymbol{\mathcal{H}}$ can be a general CD mapping and the noise can have an arbitrary covariance matrix $\mathbf{K_n}$, so long as both of these quantities are known. Methods of parameter estimation to be developed in Chap. 13 can then be used to estimate the spectral parameters. Thus a stationary or quasistationary texture field can be imaged through a shift-variant imaging system and have nonstationary noise added to it, yet the parameters describing the spectrum of the texture field can still be estimated.

Gray-level statistics When the correlation properties are not sufficient to characterize a texture, we can also use the single-point PDF $\mathrm{pr}[f(\mathbf{r})]$. For a stationary texture, this density is independent of $\mathbf{r}$, and we might want to generate samples of the texture with this density and some specified autocorrelation function or power spectral density. We shall sketch an iterative algorithm for this purpose.

The algorithm begins by filtering white noise to obtain several samples with the requisite power spectrum. It is probably valid to invoke the central-limit theorem on the filter output since the filter will serve to add up many independent samples of the white noise, so the single-point PDF on the filter output is probably Gaussian, but in any case we can estimate the PDF from the average gray-level histogram of the samples. At this stage we can perform a process known as *histogram equalization*, a pointwise nonlinear transformation that changes the gray-level distribution as described in Sec. C.3.1, and the form of the transformation can be chosen to yield the required PDF. This transformation changes the power spectrum in a complicated way, and it is necessary to estimate the new spectrum from the samples. From the new spectrum, we can devise a new filter to match the current spectrum to the required one, but this changes the PDF so a new histogram-equalization step is needed. The process is then repeated iteratively. Each iteration is a projection onto convex sets, as discussed in detail in Sec. 15.4.5, and convergence can be proven by use of a theorem quoted there. The result is a set of samples that have both the specified power spectrum and the specified single-point PDF.

Texture synthesis with wavelet channels It has been found (Bergen and Adelson, 1991; Chubb and Landy, 1991) that textures that give similar gray-level histograms through a series of wavelet filters appear similar to a human observer. Heeger and Bergen (1995) and Rolland and co-workers (Rolland and Strickland, 1997; Rolland *et al.*, 1998; Rolland, 2000) have used this observation to develop algorithms for synthesizing textures.

The Rolland group uses a digital image of a reference texture and synthesizes additional sample textures of similar visual appearance. The reference texture is decomposed into subbands by means of a discrete wavelet transform (see Sec. 5.3.3). This transform is invertible, so the original reference texture can be recovered by the inverse transform. The stochastic model, however, is that the texture can be characterized by means of gray-level histograms for each subband, basically a histogram estimate of the univariate PDFs for the output of each wavelet filter. In principle, multiple reference images could be used to improve this estimate, but the Rolland algorithm uses just one and implicitly assumes ergodicity.

To synthesize a sample texture, a discrete white noise field is generated, and it is also passed through the same discrete wavelet transform. The histogram of each filter output is computed, just as for the reference texture. A nonlinear point operation is applied in each subband to convert the histograms of the transformed white noise to histograms that match those of the reference texture. An inverse wavelet transform then yields the synthesized texture. The visual correspondence between the reference texture and the synthesized textures is striking, yet all of the synthesized textures are statistically independent since independent noise fields are used.

Multiple filters and maximum entropy The method of Heeger and Bergen permits the synthesis of textures from one or more training images, but it does not give a probability model for the synthesized images. This gap was filled by Zhu *et al.* (1998), whose work can be seen as a combination of wavelet-based texture synthesis and independent components analysis. Rather than restricting attention to some chosen set of wavelets, as in Heeger's method, Zhu *et al.* use a large library of linear filters and compute marginal histograms of the filter outputs for some training set of images (which may consist of just a single image plus an ergodicity assumption). They then use the principle of maximum entropy to construct a multivariate distribution that agrees with the marginals estimated from training data.

The rationale for maximum entropy is the one mentioned in Sec. 8.4.3: maximum-entropy densities are maximally noncommittal and do not attempt to represent information not available empirically. According to Zhu, the maximum-entropy density is the "purest fusion" of the empirical marginals.

Suppose we have a set of linear operators $\boldsymbol{\mathcal{L}}^{(j)}$ in object space, with the output of the j^{th} operator given by $q^{(j)}(\mathbf{r}) = [\boldsymbol{\mathcal{L}}^{(j)}\mathbf{f}](\mathbf{r})$. In the abstract notation of Sec. 8.2.2, the single-point marginal density on the output can be written as

$$\mathrm{pr}\left[q^{(j)}(\mathbf{r})\right] = \int d\mathbf{f}\;\mathrm{pr}\left[q^{(j)}(\mathbf{r})\middle|\mathbf{f}\right]\mathrm{pr}(\mathbf{f})\,. \tag{8.297}$$

But the linear operator is deterministic, so $q^{(j)}(\mathbf{r})$ is known exactly once $\mathbf{f}$ is specified, and we can write

$$\mathrm{pr}\left[q^{(j)}(\mathbf{r})\right] = \int d\mathbf{f}\;\delta\Big\{q^{(j)}(\mathbf{r}) - [\boldsymbol{\mathcal{L}}^{(j)}\mathbf{f}](\mathbf{r})\Big\}\,\mathrm{pr}(\mathbf{f})\,, \tag{8.298}$$

where $\delta\{q^{(j)}(\mathbf{r}) - [\boldsymbol{\mathcal{L}}^{(j)}\mathbf{f}](\mathbf{r})\}$ is simply a 1D delta function. Comparing this expression to (4.173), we see that the single-point marginal on the filter output is a Radon-transform projection of the object density $\mathrm{pr}(\mathbf{f})$, where $\mathbf{f}$ here corresponds to the position vector $\mathbf{r}$ in (4.173), and choice of the linear operator here corresponds to the projection direction $\hat{\mathbf{n}}$ in (4.173).

Now suppose we have a set of training "objects" (either good computer simulations or images from a high-resolution, low-noise imaging system as discussed in Sec. 8.4.1) from which we can form a histogram estimate of $\mathrm{pr}\left[q^{(j)}(\mathbf{r})\right]$. If we denote this histogram, defined as in (8.262), by $\widehat{\mathrm{pr}}_{q^{(j)}(\mathbf{r})}\left[q^{(j)}(\mathbf{r})\right]$, then we can pose the maximum-entropy density-estimation problem as

$$-\int d\mathbf{f}\;\mathrm{pr}(\mathbf{f})\ln[\mathrm{pr}(\mathbf{f})] = \max\,, \tag{8.299}$$

subject to the constraints of normalization,

$$\int d\mathbf{f}\;\mathrm{pr}(\mathbf{f}) = 1\,, \tag{8.300}$$

and agreement with the empirical histograms,

$$\widehat{\mathrm{pr}}_{q^{(j)}(\mathbf{r})}(z) = \int d\mathbf{f}\;\delta\Big\{z - [\boldsymbol{\mathcal{L}}^{(j)}\mathbf{f}](\mathbf{r})\Big\}\,\mathrm{pr}(\mathbf{f})\,. \tag{8.301}$$

If we assume stationarity, at least over some restricted region, then the histogram should be the same for all positions, and we can drop the argument $\mathbf{r}$ on the subscript, but we still have to satisfy the constraint at all $\mathbf{r}$. In practice, the matching will be done for a discrete set of points $\mathbf{r}_i$, usually on a pixel grid.

This problem can be solved by the method of Lagrange multipliers, just as in (8.288) *ff.*, but now we have an infinite number of constraints! For each operator $\boldsymbol{\mathcal{L}}^{(j)}$, we must satisfy (8.301) for all $\mathbf{r}$ and all z. We thus have a continuum of unknown Lagrange multipliers, which we can express as an unknown function $\Phi^{(j)}\{z\}$. With this view, the general form of the maximum-entropy object density turns out to be [see Zhu *et al.* (1998) for details]

$$\begin{aligned}\mathrm{pr}(\mathbf{f}) &= \frac{1}{Z}\exp\Big\{-\sum_i\sum_j\int dz\ \Phi^{(j)}\{z\}\,\delta\Big\{z-\Big[\boldsymbol{\mathcal{L}}^{(j)}\mathbf{f}\Big](\mathbf{r}_i)\Big\}\Big\}\\ &= \frac{1}{Z}\exp\Big\{-\sum_i\sum_j\Phi^{(j)}\Big(\Big[\boldsymbol{\mathcal{L}}^{(j)}\mathbf{f}\Big](\mathbf{r}_i)\Big)\Big\}\,, \end{aligned} \tag{8.302}$$

where Z is a normalizing constant.

The problem is not yet solved since we still have to find the functionals $\Phi^{(j)}$ such that the constraints are satisfied. Zhu *et al.* propose an iterative algorithm for this purpose.

One remaining question is how to choose the operators $\boldsymbol{\mathcal{L}}^{(j)}$ in the first place. Since stationarity is probably required to make this whole approach computationally feasible, it is natural to choose the operators as LSIV filters, but another consideration is independence. The maximum-entropy estimate in (8.302) shows that the filter outputs are statistically independent, even if this is not the case in reality. Zhu *et al.* propose use of a large library of filters and an iterative algorithm to select a subset of them that optimize a measure of independence, as in ICA, and Zhang (2001) suggests a Metropolis algorithm.

Parametric descriptions of the marginals The filters chosen in the Zhu approach (or discovered in ICA) are mostly band-pass filters (though Zhu includes a low-pass filter as well). As discussed in Sec. 8.4.3 and illustrated in Fig. 8.7, the outputs of band-pass filters tend to have simple cuspy shapes with long, kurtotic tails. Empirically, we can describe these marginals by simple analytical forms such as Laplacian or Levy densities, with only one or two free parameters per filter output.

This observation suggests an alternative to the Zhu method: instead of trying to choose the functions $\Phi^{(j)}\{z\}$ to match the empirical marginal histograms, we can directly estimate the free parameters in the assumed analytical densities (Kupinski *et al.*, 2003c).

Lumpy backgrounds Another way to generate images (or simulated objects) with specified correlation properties and controllable gray-level statistics is the *lumpy background*, introduced by Rolland and Barrett (1992). In this method, spatial elements, called *lumps* and denoted $l(\mathbf{r})$, are randomly distributed over some area, so the distribution has the form

$$f(\mathbf{r}) = \sum_{n=1}^{N} l(\mathbf{r}-\mathbf{r}_n)\,. \tag{8.303}$$

A common choice for $l(\mathbf{r})$ is a Gaussian spatial distribution,

$$l(\mathbf{r}) = A\exp\left(-\frac{r^2}{2s^2}\right)\,. \tag{8.304}$$

The positions $\mathbf{r}_n$ and possibly also the total number of lumps N are random variables.

One important special case is where N is a Poisson random variable; the mathematical tools for analyzing this case will be developed in Sec. 11.3.9, and the characteristic functional for the random process (8.303) will be derived in Sec. 11.3.10.

As we shall see there, $f(\mathbf{r})$ is a stationary random process if the positions $\mathbf{r}_n$ are uniformly distributed over some area, and the statistical autocorrelation function turns out to be just the autocorrelation integral of the lump profile [see (11.140)].

If N is large and the lump positions are statistically independent, the single-point PDF of a lumpy background approaches a Gaussian by the central-limit theorem. In this limit, the details of the lump profile are irrelevant, and the resulting functions are indistinguishable from ones obtained by filtering white, Gaussian noise. If N is small, on the other hand, then the lump profile controls the single-point PDF as well as the correlation properties; for more details, see Sec. 11.3.10.

More general lumpy backgrounds As originally defined by Rolland, the lump profile $l(\mathbf{r})$ in (8.303) is a nonrandom function; the only randomness is in the lump location $\mathbf{r}_n$. To allow more freedom in synthesizing lumpy backgrounds, we can let the lump profile also be random. For example, the amplitude or the width of each blob could vary according to some specified probability law.

One very useful variant of the simple lumpy background is the *clustered lumpy background*, suggested by Bochud *et al.* (1999a), where a cluster of identical blobs forms a *superblob*, and the final model is a superposition of superblobs. With this scheme, (8.303) becomes

$$f(\mathbf{r}) = \sum_{k=1}^{N_s} \sum_{n=1}^{N_k} l_k(\mathbf{r} - \mathbf{r}_{nk} - \mathbf{R}_k) \,, \tag{8.305}$$

where N_s is the number of superblobs, N_k is the number of blobs within the k^{th} superblob, $\mathbf{R}_k$ is the center of the k^{th} superblob, $\mathbf{r}_{nk}$ is the center of the n^{th} blob within the k^{th} superblob, and $l_k(\mathbf{r})$ is the random lump profile associated with that superblob. It is useful to make N_k and N_s Poisson random variables, so we must wait until Chap. 11 to analyze the statistics of (8.305).

Bochud *et al.* chose elongated Gaussians for the lump profiles and used their orientation as the random parameter in $l_k(\mathbf{r})$. With this simple model they were able to synthesize images strikingly similar to clinical mammograms.

Two-point densities As we discussed in Sec. 8.4.2, two-point PDFs of the form $\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)]$ can be an important part of the stochastic description of objects in general, and they are particularly attractive for stationary random processes such as textures. For purposes of stochastic modeling, we can estimate the two-point PDF from the empirical co-occurrence statistics of one or a few images if we assume ergodicity. It was suggested by Julesz (1962) that textures with similar co-occurrence statistics would appear similar, though psychophysical studies have shown that higher-order statistics do have at least some effect on human texture perception (Diaconis and Freedman, 1981).

The use of co-occurrence statistics for synthesis of realistic textures should be distinguished from their use in texture discrimination or segmentation. In the latter application, the goal is to describe the texture pattern within a spatial region by a few features with good discriminatory power, and it is common to reduce pixel values in the region to a co-occurrence matrix and to derive the features from that matrix. There is no need to make any argument about ergodicity or stationarity in this application; if the features are useful in discriminating one region from another or classifying regions, that is justification enough.

Quasistationary textures Most of the discussion above has concentrated on textures as stationary random processes. If exact stationarity is not a good assumption, we may want to model a texture as quasistationary, and in this case the stochastic Wigner distribution function defined in Sec. 8.2.5 is a useful tool. In particular, if the quasistationary form (8.142) is valid, we can estimate the two factors $b(\mathbf{r}_0)$ and $A(\boldsymbol{\rho})$ separately from samples. If we want to generate sample textures with a stochastic Wigner distribution specified by (8.142), we can use a lumpy background with a spatially variable lump density (mean number of lumps per unit area) given by $b(\mathbf{r}_0)$. For more discussion on the statistics of lumpy backgrounds, see Sec. 11.3.10.

Sometimes the pattern we want to synthesize is stationary within prespecified boundaries. For example, we may want to simulate statistically independent sample functions of an abdominal section of the body in order to study image quality in computed tomography (CT). We can start with a good anatomical model, obtained perhaps by manual or automated segmentation of a single reference CT image, and we can identify specific organs such as liver and spleen within this image (Zubal *et al.*, 1994). Then any of the methods described above can be used to characterize the texture within each organ and to generate sample functions consistent with this characterization. These sample functions can then be placed within the specified organ boundaries, and the procedure can be repeated as many times as needed to get a large number of simulated abdomens. These simulations can be regarded as object representations rather than images since the organ boundaries will be sharp and the textures may contain very high spatial frequencies.

Random shapes In addition to simulating random textures within a region, we may wish to make the shape itself random. Simulating a shape usually means adopting some parameterized description of the shape and choosing the parameters. Some simple approaches to describing shapes mathematically were discussed briefly in Sec. 7.1.6. One approach, used for example by Cargill (1989) to describe the human liver, is to specify the distance R from some internal reference point to the boundary as a function of polar angles θ and ϕ. If the surface of the object is smooth, an expansion of $R(\theta, \phi)$ in spherical harmonics can be terminated with relatively few terms (~ 100 in Cargill's work), and the coefficients in this expansion are the desired parametric representation of the liver. This general approach is applicable to any 3D shape in which a reference point can be found for which $R(\theta, \phi)$ is unique; it is not necessary that the shape be convex, though convexity avoids the necessity of searching for a suitable reference point.

Another general approach, also mentioned briefly in Sec. 7.1.6, is to express the shape as a geometric transformation of a given reference shape. Affine or non-affine transformations can be used, and the parameters of the transformation are then the shape descriptors.

After establishing a parametric description of shape, the next step in shape simulation is to find the PDF on the parameters, for example by analyzing real shapes. One common approach is to compute a sample mean and sample covariance matrix on a set of measured parameters and, in effect, to assume that the PDF is multivariate normal with this mean and covariance. If there are many parameters, it can be advantageous to use principal components analysis or PCA (see Sec. 8.4.1) and retain only components corresponding to a few of the eigenvectors of the sample covariance with the largest eigenvalues. The eigenvectors themselves are sets of shape parameters, and the shapes associated with them are often called *eigenshapes.* It must be kept in mind, however, that these eigenshapes are characteristics of both the particular shape description used and the experimental data set from which the parameters were derived.

However the PDF on the shape parameters is formulated, samples drawn from it can be used to synthesize new shapes consistent with the estimated PDF, and these random shapes can then be used in image-quality studies and many other investigations. For an example of these procedures, see Duta *et al.* (1999), and for general mathematical treatments of statistical shape analysis, see Small (1996), Dryden and Mardia (1998) and Kendall *et al.* (1999).

8.4.5 Signals and backgrounds

In many imaging situations, we do not have equal interest in all parts of the scene. In aerial reconnaissance, for example, we are relatively uninterested in trees and bushes, but we would be extraordinarily interested in a military vehicle that might be hiding in the bushes. Similarly, in an abdominal MRI scan, we have little interest in the myriad features of normal anatomy, but we are much more interested in a small nodule that might turn out to be malignant. As a very general term, we can call an object of interest, that may or may not be present in a given scene, a *signal.* The remainder of the scene can be called *background* or (especially in the radar literature) *clutter.* In Chap. 13 we shall discuss in detail methods of detecting signals, or distinguishing between different signals, but here we introduce the topic by discussing stochastic models for objects with and without signals.

Additive signals Perhaps surprisingly, it entails no loss of generality to decompose an object into a simple sum of signal and background components:

$$f(\mathbf{r}) = f_s(\mathbf{r}) + f_b(\mathbf{r})\,. \tag{8.306}$$

Once we have defined the portion of the object that we regard as signal and denoted it as $f_s(\mathbf{r})$, then the background $f_b(\mathbf{r})$ is just *defined* as $f(\mathbf{r}) - f_s(\mathbf{r})$.

This does not say that $f_b(\mathbf{r})$ is the same as $f(\mathbf{r})$ would be in the absence of the signal, though in fact it may be. In nuclear medicine for example, a tumor is often manifest by an increased uptake of some tumor-seeking radiopharmaceutical, so it is natural to simply add the tumor distribution $f_s(\mathbf{r})$ to the distribution $f_b(\mathbf{r})$ in normal tissue. If both $f_s(\mathbf{r})$ and $f_b(\mathbf{r})$ are sample functions of random processes, then it may be reasonable to take the two processes as statistically independent.

In optical imaging, on the other hand, objects are opaque, so a signal of interest may obscure the background behind it. For purposes of describing the response of an imaging system, the object $f(\mathbf{r})$ is either $f_b(\mathbf{r})$ or $f_s(\mathbf{r})$, not their sum. Nevertheless, we can still use an additive model if the statistical dependence of the two processes is taken into account.

Nonrandom signals The simplest model for a signal on a background is one where the signal function is completely specified and the only randomness is whether or not it is present. This model is often called SKE (signal known exactly). As we shall see in Chap. 13, it is an excellent starting point for discussing signal detection and image quality.

If we adopt the SKE model and assume that the signal is just added to the background rather than obscuring it, then the signal and background are statistically independent. With or without a signal, the PDF on $f(\mathbf{r})$ is fully determined by the PDF on $f_b(\mathbf{r})$ since that is the only random process in the problem. In the absence of a signal, we can write the univariate density on $f(\mathbf{r})$ as

$$\text{pr}[f(\mathbf{r})|\text{signal absent}] = \text{pr}_b[f_b(\mathbf{r})] \,. \tag{8.307}$$

We have added the subscript b to indicate that $\text{pr}_b[f_b(\mathbf{r})]$ is specifically the PDF on the background; the notation is redundant here since the same information is conveyed by the subscript on $f_b(\mathbf{r})$, but its usefulness will become apparent in a moment.

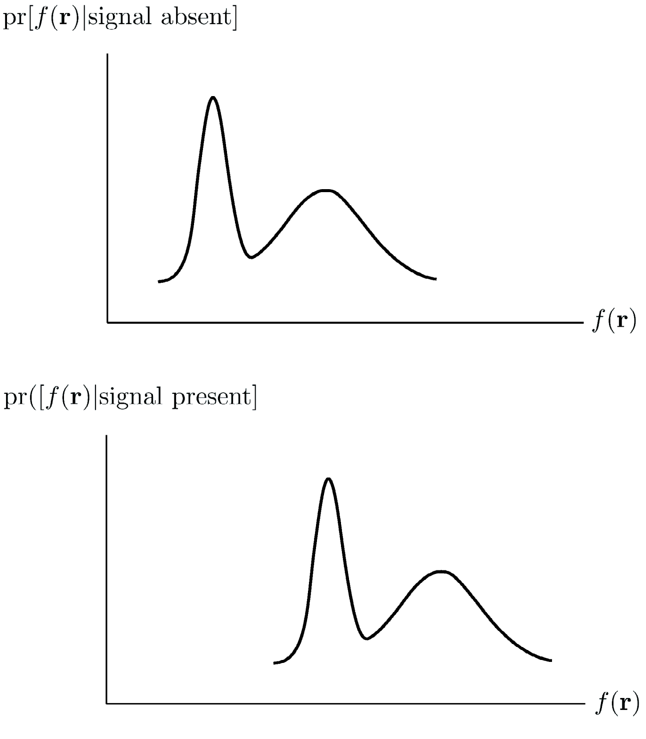
Because of the assumed statistical independence, the form of the PDF for $f_b(\mathbf{r})$ is still the same with a signal present, but to relate it to the PDF on $f(\mathbf{r})$ we must rewrite (8.306) as

$$f_b(\mathbf{r}) = f(\mathbf{r}) - f_s(\mathbf{r}) \,. \tag{8.308}$$

We then have

$$\text{pr}[f(\mathbf{r})|\text{signal present}] = \text{pr}_b[f(\mathbf{r}) - f_s(\mathbf{r})] \,. \tag{8.309}$$

Now we see the need for the subscript: the 1D PDF $\text{pr}_b[f_b(\mathbf{r})]$ is merely shifted along the axis by the presence of a nonrandom signal (see Fig. 8.9), and the functional form is unchanged.



$\text{pr}[f(\mathbf{r})|\text{signal absent}]$

$\text{pr}([f(\mathbf{r})|\text{signal present}]$

Fig. 8.9 Effect on the univariate PDF of adding a nonrandom signal to a random background.

This discussion has specifically dealt with univariate densities at a single point, but it is easy to extend it to an arbitrary number of points or to general Hilbert-space vectors representing object, signal and background. Abstractly, we can write

$$\mathrm{pr}(\mathbf{f}\,|\text{signal absent}) = \mathrm{pr}_b(\mathbf{f}_b)\,; \tag{8.310}$$

$$\mathrm{pr}(\mathbf{f}\,|\text{signal present}) = \mathrm{pr}_b(\mathbf{f} - \mathbf{f}_s)\,. \tag{8.311}$$

These densities can be interpreted as PDFs on coefficient vectors like $\boldsymbol{\alpha}$ or as multipoint densities. In fact, (8.308) and (8.309) follow from (8.310) and (8.311) just by regarding $f(\mathbf{r})$ as a component of $\mathbf{f}$ (and similarly for $\mathbf{f}_b$ and $\mathbf{f}_s$) and taking marginals on both sides of (8.310) and (8.311). Whatever space we are working in, addition of a nonrandom signal merely shifts the background PDF.

Parametric signal models Sometimes a signal is not known exactly but can be described by a function with a small number of unknown parameters. For example, in nuclear medicine a tumor might be well modeled as a sphere with random location, size and uptake of a radiopharmaceutical. Similarly, in astronomy a pulsar could be modeled as a time-varying point source, where the random parameters are its coordinates in the sky and the amplitude and period of the pulsation.

In these cases we do not need an infinite-dimensional PDF like $\mathrm{pr}(\mathbf{f}_s)$ to describe the signal; if the signal is fully specified by L parameters $\{\theta_{s\ell}, \ell = 1, ..., L\}$, all we need is the L-variate PDF $\mathrm{pr}(\{\theta_{s\ell}\})$. The signal parameters can also be arranged into an $L \times 1$ vector $\boldsymbol{\theta}_s$, so we need the PDF $\mathrm{pr}(\boldsymbol{\theta}_s)$ in order to describe the signal fully.

With a signal described parametrically, the object PDF is given by

$$\mathrm{pr}(\mathbf{f}\,|\text{signal present}) = \int_\infty d^L\theta \;\mathrm{pr}(\mathbf{f}\,|\text{ signal present, } \boldsymbol{\theta}_s)\,\mathrm{pr}(\boldsymbol{\theta}_s)\,. \tag{8.312}$$

The conditional density $\mathrm{pr}(\mathbf{f}\,|\text{ signal present, } \boldsymbol{\theta}_s)$ is just the density for an SKE problem; if we condition on a set of parameters that completely specify the signal, then the signal *is* known exactly. If the signal and background are statistically independent, then this conditional density is given by (8.311), and we have

$$\mathrm{pr}(\mathbf{f}\,|\text{signal present}) = \int_\infty d^L\theta_s \;\mathrm{pr}_b\left[\mathbf{f} - \mathbf{f}_s(\boldsymbol{\theta}_s)\right]\mathrm{pr}(\boldsymbol{\theta}_s)\,. \tag{8.313}$$

The object PDF is now a weighted average of shifted background PDFs.

Obscuring signals If point $\mathbf{r}$ lies within the signal and the signal obscures the background, then $f_b(\mathbf{r})$ can take on only the value zero at this point. Since $f(\mathbf{r})$ is then identical to $f_s(\mathbf{r})$, the univariate density on $f(\mathbf{r})$ for a nonrandom signal is given by

$$\mathrm{pr}[f(\mathbf{r})|\text{signal present at } \mathbf{r}] = \delta[f(\mathbf{r}) - f_s(\mathbf{r})]\,. \tag{8.314}$$

If the signal is absent, or if it is present in the object but not at point $\mathbf{r}$, then (8.307) still holds.

Multipoint densities can be formulated similarly. For example, if a nonrandom signal is present at $\mathbf{r}_1$ but not at $\mathbf{r}_2$, the two-point conditional PDF is

$$\mathrm{pr}[f(\mathbf{r}_1), f(\mathbf{r}_2)|\text{signal present at } \mathbf{r}_1\text{, absent at } \mathbf{r}_2\,] = \delta[f(\mathbf{r}_1) - f_s(\mathbf{r}_1)]\,\mathrm{pr}_b[f(\mathbf{r}_2)]\,. \tag{8.315}$$

The univariate marginals of (8.315) are consistent with (8.314) and (8.307).

The PDFs specified by (8.314) and (8.315) can be difficult to work with, especially when we extend the discussion to random signals. It is often preferable to work in terms of the expansion coefficients $\{\alpha_n\}$. Consider a nonrandom, obscuring signal with support $\mathbf{S}_s$; that is, the signal obscures the background for all points $\mathbf{r}$ in the region $\mathbf{S}_s$. For an orthonormal basis, the coefficient α_n is given by

$$\alpha_n = \int_{\mathbf{S}_f} d^q r \ \psi_n^*(\mathbf{r}) \, f(\mathbf{r}) \, , \tag{8.316}$$

where $\mathbf{S}_f$ is the overall support of the object. When a signal is present, this integral can be written as

$$\alpha_n = \int_{\mathbf{S}_s} d^q r \ \psi_n^*(\mathbf{r}) \, f_s(\mathbf{r}) + \int_{\mathbf{S}_{sc}} d^q r \ \psi_n^*(\mathbf{r}) \, f_b(\mathbf{r}) \, , \tag{8.317}$$

where $\mathbf{S}_{sc}$ is the complement of $\mathbf{S}_s$, *i.e.*, the set of points in $\mathbf{S}_f$ but not in $\mathbf{S}_s$.

We can think of the first integral in (8.317) as the n^{th} component of an infinite vector $\boldsymbol{\alpha}_s$ describing the signal in the basis $\{\psi_n\}$; since the signal is nonrandom, $\boldsymbol{\alpha}_s$ is nonrandom. The second integral would be the n^{th} expansion coefficient for the background except that we have excluded the region $\mathbf{S}_s$ from the range of integration. Nevertheless, we can think of that integral as the n^{th} component of a random vector which we can denote as $\boldsymbol{\alpha}_b$, and we can write, without approximation,

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}_s + \boldsymbol{\alpha}_b \, , \qquad \text{(signal present)} \, . \tag{8.318}$$

If the signal is absent, then the support of the background is the same as the object support, and we can write

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}_b \, , \qquad \text{(signal absent)} \, , \tag{8.319}$$

where $\boldsymbol{\alpha}_b$ is now computed via integration over all of $\mathbf{S}_f$.

Because of the different regions of integration, the statistics of $\boldsymbol{\alpha}_b$ will, in general, depend on whether or not the signal is present. There is, however, one interesting situation in which we might assume that $\boldsymbol{\alpha}_b$ is independent of the signal. Suppose we have a spatially compact signal but a spatially extended basis function, such as a Fourier basis function (see Sec. 7.1.2). In that case, deletion of a small region may not change the value of the integral very much, so it might be a good approximation to say that $\boldsymbol{\alpha}_b$ is the same with and without the obscuring signal. If that assumption is valid, then we are back to an additive model with a signal-independent background, at least in this basis. If, on the other hand, deletion of the signal support does change the integral significantly, we can still use the additive form (8.318), but we have to use a different PDF on $\boldsymbol{\alpha}_b$ for signal present and signal absent.

8.5 STOCHASTIC MODELS FOR IMAGES

Having just discussed various stochastic models for objects, we turn now to images. In keeping with our emphasis on digital imaging, we consider only CD systems here, and for simplicity we assume they are linear. Our objective will be to characterize an ensemble of such images by its mean vector and covariance matrix and, where possible, a multivariate probability density function.

8.5.1 Linear systems

In the absence of noise, we defined a linear imaging system as one for which the image was a linear functional of the object; with noise, a linear imaging system can be defined as one for which the *average* image, obtained after many repeated images of the same object, is a linear functional of the object. If we denote this mean image by $\overline{\mathbf{g}}(\mathbf{f})$, then for any linear system we can write

$$\overline{\mathbf{g}}(\mathbf{f}) = \boldsymbol{\mathcal{H}}\mathbf{f}\,, \tag{8.320}$$

where $\boldsymbol{\mathcal{H}}$ is a linear operator acting on the object $\mathbf{f}$.

Specifically for the case of digital imaging of an object function, we know from Sec. 7.3.1 that the most general way to write the linear mapping is

$$\overline{g}_m(\mathbf{f}) = \int_{\mathbf{S}_f} d^q r\; h_m(\mathbf{r}) f(\mathbf{r})\,, \qquad m = 1, ..., M\,. \tag{8.321}$$

Except for the overbar and the explicit argument $\mathbf{f}$, this equation is identical to (7.225). We emphasize that the average implied by this overbar is for repeated images of a single object.

To get an expression for the actual random image, we can define an $M \times 1$ noise vector $\mathbf{n}$ by

$$\mathbf{n} \equiv \mathbf{g} - \overline{\mathbf{g}}(\mathbf{f}) = \mathbf{g} - \boldsymbol{\mathcal{H}}\mathbf{f}\,. \tag{8.322}$$

Thus, we have

$$\mathbf{g} = \boldsymbol{\mathcal{H}}\mathbf{f} + \mathbf{n}\,. \tag{8.323}$$

This is the fundamental equation describing noisy, digital imaging of real objects. Now we must understand the statistical properties of $\mathbf{g}$, both for a particular object $\mathbf{f}$ and when a random ensemble of objects is considered.

8.5.2 Conditional statistics for a single object

Conditional density If each component of $\mathbf{g}$ is a continuous random variable, we can denote the conditional probability density function (for a particular object) by $\mathrm{pr}(\mathbf{g}|\mathbf{f})$. If each component of $\mathbf{g}$ can take on only discrete values, we should use the conditional probability $\mathrm{Pr}(\mathbf{g}|\mathbf{f})$, but to avoid considering these two cases in parallel, we shall use the lower-case $\mathrm{pr}(\mathbf{g}|\mathbf{f})$ in both cases, understanding it as a probability density function or probability as needed. Specific forms for $\mathrm{pr}(\mathbf{g}|\mathbf{f})$ will be given later, especially in Chaps. 11 and 12. As we shall see there, independent Poisson models are usually valid when photon-counting detectors are used, and multivariate normal models are valid with most other detectors.

Even without specific models for the detector statistics, we can make some general statements about $\mathrm{pr}(\mathbf{g}|\mathbf{f})$. For one thing, we know that $\mathbf{f}$ affects the data only through the system operator $\boldsymbol{\mathcal{H}}$, so

$$\mathrm{pr}(\mathbf{g}|\mathbf{f}) = \mathrm{pr}(\mathbf{g}|\boldsymbol{\mathcal{H}}\mathbf{f}) = \mathrm{pr}(\mathbf{g}|\boldsymbol{\mathcal{H}}\mathbf{f}_{meas})\,. \tag{8.324}$$

Thus only the measurement component of the object affects the statistics of the image.

Furthermore, for a given $\mathbf{f}$, $\boldsymbol{\mathcal{H}}\mathbf{f}$ is not a random variable, so

$$\mathrm{pr}(\mathbf{g}|\mathbf{f}) = \mathrm{pr}_{\mathbf{n}}(\mathbf{g} - \boldsymbol{\mathcal{H}}\mathbf{f}\,|\boldsymbol{\mathcal{H}}\mathbf{f})\,, \tag{8.325}$$

where $\mathrm{pr}_{\mathbf{n}}(\mathbf{n}|\overline{\mathbf{g}})$ is the PDF on the noise vector[14] given some mean value for the detector output. If this density is independent of $\overline{\mathbf{g}}$, then we say that the noise is *object-independent* and write

$$\mathrm{pr}(\mathbf{g}|\mathbf{f}) = \mathrm{pr}_{\mathbf{n}}(\mathbf{g} - \boldsymbol{\mathcal{H}}\mathbf{f}) . \tag{8.326}$$

In this case, therefore, the conditional density on $\mathbf{g}$ is just a displaced version of the density on the noise. As we shall see in more detail in later chapters, this object-independent model is often valid for electronic and other excess noise in detectors.

A related approximation is that the noise is object-dependent but *signal-independent.* When we divide an object into signal and background, as in (8.306), it may turn out that the signal is weak compared to the background, and sometimes we can write $\mathrm{pr}(\mathbf{g}|\mathbf{f}) = \mathrm{pr}(\mathbf{g}|\mathbf{f}_b + \mathbf{f}_s) \approx \mathrm{pr}(\mathbf{g}|\mathbf{f}_b)$. This may be a good approximation with photon-counting detectors in low-contrast situations where all components of $\boldsymbol{\mathcal{H}}\mathbf{f}$ are approximately equal.

Another assumption that is often justified in practice is that the components of $\mathbf{n}$ are statistically independent for a fixed object. With discrete arrays of photodiodes, for example, the electronic noise in one element is often statistically independent of noise in all other elements, and we shall see in Chap. 11 that photon-counting detectors viewing a Poisson source almost always yield statistically independent measurements. When this assumption is valid, we have

$$\mathrm{pr}(\mathbf{g}|\mathbf{f}) = \prod_{m=1}^{M} \mathrm{pr}(g_m|\mathbf{f}) . \tag{8.327}$$

Conditional mean and covariance We can also make some general statements about conditional means and covariances. We know already that the conditional mean of $\mathbf{g}$ is

$$\mathrm{E}\{\mathbf{g}|\mathbf{f}\} \equiv \overline{\mathbf{g}} = \boldsymbol{\mathcal{H}}\mathbf{f} , \tag{8.328}$$

from which it follows at once that

$$\mathrm{E}\{\mathbf{n}|\mathbf{f}\} = 0 . \tag{8.329}$$

Thus we can always regard the noise vector as zero-mean.

Since we are conditioning on $\mathbf{f}$ and hence $\boldsymbol{\mathcal{H}}\mathbf{f}$ is not a random variable, the conditional covariance of $\mathbf{g}$ is the same as the covariance of $\mathbf{n}$; notationally, we write

$$\mathbf{K}_{\mathbf{g}|\mathbf{f}} = \mathbf{K}_{\mathbf{n}} , \tag{8.330}$$

but we must allow for the possibility that $\mathbf{K}_{\mathbf{n}}$ depends on $\mathbf{f}$ (in the Poisson case, for example).

8.5.3 Effects of object randomness

Next we examine the image statistics in the case where the object is random. In frequentist terms, we can consider a large number of images, each with a different object drawn from some ensemble. Our knowledge of the object statistics is given by a stochastic model such as those considered in Sec. 8.4.

---

[14] Recall that we add subscripts to PDFs only when the random variable is not obvious from the argument.

Overall density Formally, we can write the overall probability density as

$$\mathrm{pr}(\mathbf{g}) = \int_{\mathbb{U}} d\mathbf{f}\ \mathrm{pr}(\mathbf{g}|\mathbf{f})\,\mathrm{pr}(\mathbf{f})\,. \tag{8.331}$$

In principle, this integral runs over the entire infinite-dimensional object space, but from (8.324) we know that only the measurement subspace contributes. The dimensionality of this subspace is R, the rank of the operator $\boldsymbol{\mathcal{H}}$, so really only R components are important. If we expand $\mathbf{f}_{meas}$ in some suitable basis for measurement space as in (7.251), with an $R \times 1$ coefficient vector $\boldsymbol{\alpha}$, then the integral can be written as

$$\mathrm{pr}(\mathbf{g}) = \int_{\infty} d^R\alpha\ \mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha})\,\mathrm{pr}(\boldsymbol{\alpha})\,. \tag{8.332}$$

Depending on the choice of basis for measurement space, there is some matrix $\mathbf{H}_0$ that exactly maps the coefficients $\boldsymbol{\alpha}$ to $\boldsymbol{\mathcal{H}}\mathbf{f}$ (see Sec. 7.4.3), so we can write

$$\mathrm{pr}(\mathbf{g}) = \int_{\infty} d^R\alpha\ \mathrm{pr}(\mathbf{g}|\mathbf{H}_0\boldsymbol{\alpha})\,\mathrm{pr}(\boldsymbol{\alpha})\,. \tag{8.333}$$

Derivation of the form of $\mathbf{H}_0$ for the specific case of expansion in natural pixels (Sec. 7.4.3) is an interesting exercise for the reader.

For object-independent noise as in (8.326), (8.333) takes the appealing form,

$$\mathrm{pr}(\mathbf{g}) = \int_{\infty} d^R\alpha\ \mathrm{pr}_{\mathbf{n}}(\mathbf{g} - \mathbf{H}_0\boldsymbol{\alpha})\,\mathrm{pr}(\boldsymbol{\alpha})\,. \tag{8.334}$$

This equation is not quite a convolution, but nevertheless it can be usefully transformed by Fourier methods. With characteristic functions as defined in (8.27) and some algebra similar to that used in obtaining (8.43), we can show that

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \psi_{\mathbf{n}}(\boldsymbol{\xi})\,\psi_{\boldsymbol{\alpha}}(\mathbf{H}_0^t\boldsymbol{\xi})\,. \tag{8.335}$$

Increasing the noise level decreases the width of $\psi_{\mathbf{n}}(\boldsymbol{\xi})$ in this Fourier domain, and increasing the degree of object randomness decreases the width of $\psi_{\boldsymbol{\alpha}}(\mathbf{H}_0^t\boldsymbol{\xi})$; either measure decreases the width of $\psi_{\mathbf{g}}(\boldsymbol{\xi})$ and hence increases the spread of $\mathrm{pr}(\mathbf{g})$.

For Poisson noise (8.334) and (8.335) are not valid; instead, (8.334) must be written as[15]

$$\mathrm{Pr}(\mathbf{g}) = \int_{\infty} d^R\alpha\ \mathrm{Pr}(\mathbf{g}|\mathbf{H}_0\boldsymbol{\alpha})\,\mathrm{pr}(\boldsymbol{\alpha})\,. \tag{8.336}$$

Note that we have written $\mathrm{Pr}(\mathbf{g})$ instead of $\mathrm{pr}(\mathbf{g})$ since Poisson random variables are discrete. The probability (not density) $\mathrm{Pr}(\mathbf{g}|\mathbf{H}_0\boldsymbol{\alpha})$ is just a product of univariate Poisson probabilities, where the mean of g_m is $[\mathbf{H}_0\boldsymbol{\alpha}]_m$.

The transformation of the characteristic function in the Poisson case was derived by Clarkson *et al.* (2002). They show that (8.336) is equivalent to

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \psi_{\boldsymbol{\alpha}}[\mathbf{H}_0^t\,\boldsymbol{\Gamma}(\boldsymbol{\xi})]\,, \tag{8.337}$$

---

[15] We could also have written $\mathrm{Pr}(\mathbf{g}|\mathbf{H}_0\boldsymbol{\alpha})$ in (8.336) as $\mathrm{Pr}(\mathbf{g}|\boldsymbol{\alpha})$ since $\mathbf{H}_0\boldsymbol{\alpha}$ is fully determined by $\boldsymbol{\alpha}$, but the former version is more useful when we want to write the probability as a product of Poissons; the probability for g_m is specified by a single component of $\mathbf{H}_0\boldsymbol{\alpha}$, but all components of $\boldsymbol{\alpha}$ may be required because of the matrix multiplication.

where $\mathbf{\Gamma}$ is an operator that acts independently on each component of its vector operand; it is defined such that

$$[\mathbf{\Gamma}(\boldsymbol{\xi})]_m = \frac{-1 + \exp(-2\pi i\, \xi_m)}{-2\pi i} \,. \tag{8.338}$$

Clarkson *et al.* (2002) show also that this transformation law applies when $\mathbf{H}_0$ is replaced by a CD operator $\boldsymbol{\mathcal{H}}$ and the full infinite-dimensional vector $\mathbf{f}$ is used in place of the finite-dimensional $\boldsymbol{\alpha}$. In that case the characteristic *function* for $\mathbf{g}$ is related to the characteristic *functional* for $\mathbf{f}$ by

$$\psi_{\mathbf{g}}(\boldsymbol{\xi}) = \Psi_{\mathbf{f}}[\boldsymbol{\mathcal{H}}^\dagger\, \mathbf{\Gamma}(\boldsymbol{\xi})] \,. \tag{8.339}$$

Overall mean We shall use the notation of (8.331), recognizing that the integral will be realized by expanding the measurement component of $\mathbf{f}$ in some basis and integrating over the coefficients. With this convention, we can write the overall mean image as

$$\mathrm{E}(\mathbf{g}) = \int_\infty d^M g\; \mathbf{g}\, \mathrm{pr}(\mathbf{g}) = \int_\infty d^M g\; \mathbf{g} \int_{\mathbb{U}} d\mathbf{f}\;\, \mathrm{pr}(\mathbf{g}|\mathbf{f})\, \mathrm{pr}(\mathbf{f}) \,. \tag{8.340}$$

Shuffling the integrals, we see that

$$\mathrm{E}(\mathbf{g}) = \int_{\mathbb{U}} d\mathbf{f}\;\, \mathrm{pr}(\mathbf{f}) \int_\infty d^M g\; \mathbf{g}\, \mathrm{pr}(\mathbf{g}|\mathbf{f}) \,. \tag{8.341}$$

The inner integral is the average of $\mathbf{g}$ with respect to the conditional density, which is precisely what we called $\overline{\mathbf{g}}(\mathbf{f})$ previously, so

$$\mathrm{E}(\mathbf{g}) = \int_{\mathbb{U}} d\mathbf{f}\;\, \mathrm{pr}(\mathbf{f}) \overline{\mathbf{g}}(\mathbf{f}) \,. \tag{8.342}$$

Another notation that means the same thing is

$$\langle\, \mathbf{g} \,\rangle = \langle \langle \mathbf{g} \rangle_{\mathbf{n}|\mathbf{f}} \rangle_{\mathbf{f}} \,. \tag{8.343}$$

Yet another notation denotes this overall average as $\overline{\overline{\mathbf{g}}}$, with the double overbar indicating that we have averaged over both the measurement noise and the object variability. This double average can also be seen directly in (8.340) when we recall that $\mathrm{pr}(\mathbf{g}|\mathbf{f})\, \mathrm{pr}(\mathbf{f})$ is also the joint density, $\mathrm{pr}(\mathbf{g}, \mathbf{f})$.

Overall covariance When both measurement noise and object variability are taken into account, the covariance matrix on $\mathbf{g}$ is defined (for real $\mathbf{g}$) by

$$\mathbf{K}_{\mathbf{g}} = \langle \langle\, [\mathbf{g} - \overline{\overline{\mathbf{g}}}]\; [\mathbf{g} - \overline{\overline{\mathbf{g}}}]^t \,\rangle_{\mathbf{n}|\mathbf{f}} \rangle_{\mathbf{f}} \,. \tag{8.344}$$

Adding and subtracting $\overline{\mathbf{g}}(\mathbf{f})$ in each factor gives

$$\mathbf{K}_{\mathbf{g}} = \langle \langle\, [\mathbf{g} - \overline{\mathbf{g}}(\mathbf{f}) + \overline{\mathbf{g}}(\mathbf{f}) - \overline{\overline{\mathbf{g}}}]\; [\mathbf{g} - \overline{\mathbf{g}}(\mathbf{f}) + \overline{\mathbf{g}}(\mathbf{f}) - \overline{\overline{\mathbf{g}}}]^t \,\rangle_{\mathbf{n}|\mathbf{f}} \rangle_{\mathbf{f}} \,. \tag{8.345}$$

Noting that $\overline{\mathbf{g}}(\mathbf{f}) - \overline{\overline{\mathbf{g}}}$ does not involve $\mathbf{n}$ (since it has been averaged out) and that $\langle [\mathbf{g} - \overline{\mathbf{g}}(\mathbf{f})] \rangle_{\mathbf{n}|\mathbf{f}} = 0$, we see that

$$\mathbf{K}_{\mathbf{g}} = \langle \langle [\mathbf{g} - \overline{\mathbf{g}}(\mathbf{f})][\mathbf{g} - \overline{\mathbf{g}}(\mathbf{f})]^t \rangle_{\mathbf{n}|\mathbf{f}} \rangle_{\mathbf{f}} + \langle\, [\overline{\mathbf{g}}(\mathbf{f}) - \overline{\overline{\mathbf{g}}}]\; [\overline{\mathbf{g}}(\mathbf{f}) - \overline{\overline{\mathbf{g}}}]^t \,\rangle_{\mathbf{f}} \,. \tag{8.346}$$

The first term in this expression is just the noise covariance matrix $\mathbf{K_n}$ averaged over $\mathbf{f}$ (though this average is superfluous in the case of object-independent noise); we can denote this term as $\overline{\mathbf{K}}_\mathbf{n}$. The second term has nothing to do with $\mathbf{n}$ but rather reflects the object variability as seen in the mean image; we can denote this term as $\mathbf{K}_{\overline{\mathbf{g}}}$. With this notation, we have

$$\mathbf{K_g} = \overline{\mathbf{K}}_\mathbf{n} + \mathbf{K}_{\overline{\mathbf{g}}}\,. \tag{8.347}$$

This division of the overall covariance into two terms, one representing the average noise covariance and the other representing the variation in the conditional mean, is exact and does not require any assumptions about the form of either $\mathrm{pr}(\mathbf{g}|\mathbf{f})$ or $\mathrm{pr}(\mathbf{f})$. In particular, it does not require that the noise be object-independent, and it does not require that either the noise or the object be Gaussian.

Other expressions for the object-variability term There are several alternative ways of expressing $\mathbf{K}_{\overline{\mathbf{g}}}$. First, since the object $f(\mathbf{r})$ is a sample function of a random process, we can use the autocovariance operator $\boldsymbol{\mathcal{K}}_\mathbf{f}$, *i.e.*, the integral operator with kernel $K_\mathbf{f}(\mathbf{r},\mathbf{r}')$. Since $\overline{\mathbf{g}}$ is a linear transformation of $\mathbf{f}$ by (8.320), it follows that [*cf.* (8.50) and (8.145)]

$$\mathbf{K}_{\overline{\mathbf{g}}} = \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_\mathbf{f}\boldsymbol{\mathcal{H}}^\dagger\,. \tag{8.348}$$

Similarly, if we know that $\mathbf{f}_{meas} = \mathbf{H}_0\boldsymbol{\alpha}$ as in (8.333), and if we know the covariance matrix $\mathbf{K}_{\boldsymbol{\alpha}}$, then we have

$$\mathbf{K}_{\overline{\mathbf{g}}} = \mathbf{H}_0\,\mathbf{K}_{\boldsymbol{\alpha}}\,\mathbf{H}_0^t\,. \tag{8.349}$$

Finally, if we have some approximate object representation as in (7.301) and a system matrix $\mathbf{H}$ as defined in (7.304), and we know a covariance matrix for the coefficients $\boldsymbol{\theta}$, then we can approximate $\mathbf{K}_{\overline{\mathbf{g}}}$ by

$$\mathbf{K}_{\overline{\mathbf{g}}} \approx \mathbf{H}\,\mathbf{K}_{\boldsymbol{\theta}}\,\mathbf{H}^t\,. \tag{8.350}$$

This approximation will be accurate if the image error defined in (7.329) is small for all objects in the ensemble (and, of course, if $\mathbf{K}_{\boldsymbol{\theta}}$ is accurate).

Sample averages We have written formal expressions for the overall mean and covariance as if we know the densities needed to perform the averages. In practice, we will usually know the conditional density $\mathrm{pr}(\mathbf{g}|\mathbf{f})$, since it follows from the physics of the measurement process; as we have noted, this conditional density will usually be Gaussian or Poisson. The average over objects is much more problematical in practice. In Sec. 8.4 we discussed a variety of statistical models for objects, but we saw that there were many circumstances where we could generate samples of $\mathbf{f}$ but could not develop an analytical expression for $\mathrm{pr}(\mathbf{f})$. In these circumstances we have no choice but to approximate the analytical averages with sample averages; more details on how this is done in practice will be forthcoming in Chap. 14.

8.5.4 Signals and backgrounds in image space

In Sec. 8.4.5, we divided the object into signal and background parts as in (8.306), which we can also write as

$$\mathbf{f} = \mathbf{f}_s + \mathbf{f}_b\,. \tag{8.351}$$

Now we shall look at how this division affects the image statistics.

Conditional statistics The conditional mean, for a fixed object, is still given by (8.328), but because of the assumed linearity of the operator, we can write separately that

$$\overline{\mathbf{g}}_s = \boldsymbol{\mathcal{H}}\mathbf{f}_s\,, \qquad \overline{\mathbf{g}}_b = \boldsymbol{\mathcal{H}}\mathbf{f}_b\,, \tag{8.352}$$

The conditional covariance is still given by (8.330), but for signal-dependent noise we have to assume in general that the noise covariance matrix depends on both the signal and the background. In many problems, however, we can assume that the signal is weak compared to the background, so $\mathbf{K_n}$ is approximately independent of $\mathbf{f}_s$.

The conditional density is still given by (8.325), which we can now write as

$$\mathrm{pr}(\mathbf{g}|\mathbf{f}) = \mathrm{pr}_{\mathbf{n}}(\mathbf{g} - \boldsymbol{\mathcal{H}}\mathbf{f}_s - \boldsymbol{\mathcal{H}}\mathbf{f}_b|\boldsymbol{\mathcal{H}}\mathbf{f}) \tag{8.353}$$

or, for object-independent noise,

$$\mathrm{pr}(\mathbf{g}|\mathbf{f}) = \mathrm{pr}_{\mathbf{n}}(\mathbf{g} - \boldsymbol{\mathcal{H}}\mathbf{f}_s - \boldsymbol{\mathcal{H}}\mathbf{f}_b)\,. \tag{8.354}$$

For noise that is object-dependent but signal-independent, this expression would become $\mathrm{pr}(\mathbf{g}|\mathbf{f}) = \mathrm{pr}_{\mathbf{n}}(\mathbf{g} - \boldsymbol{\mathcal{H}}\mathbf{f}_s - \boldsymbol{\mathcal{H}}\mathbf{f}_b|\boldsymbol{\mathcal{H}}\mathbf{f}_b)$.

Random background When the background $\mathbf{f}_b$ is random but the signal is not, then the overall probability density function in (8.331) becomes

$$\mathrm{pr}(\mathbf{g}) = \int_{\mathbb{U}} d\mathbf{f}_b\; \mathrm{pr}(\mathbf{g}|\mathbf{f}_b, \mathbf{f}_s)\,\mathrm{pr}(\mathbf{f}_b) \tag{8.355}$$

or, for object-independent noise,

$$\mathrm{pr}(\mathbf{g}) = \int_{\mathbb{U}} d\mathbf{f}_b\; \mathrm{pr}_{\mathbf{n}}(\mathbf{g} - \boldsymbol{\mathcal{H}}\mathbf{f}_b - \boldsymbol{\mathcal{H}}\mathbf{f}_s)\,\mathrm{pr}(\mathbf{f}_b)\,. \tag{8.356}$$

For a nonrandom signal, the overall covariance matrix is almost unchanged from before; from (8.347) and (8.348), we have

$$\mathbf{K_g} = \overline{\mathbf{K}}_{\mathbf{n}} + \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_{\mathbf{f}_b}\boldsymbol{\mathcal{H}}^\dagger\,. \tag{8.357}$$

Essentially the only change here is the subscript on $\boldsymbol{\mathcal{K}}$.

Random signals If both signal and background are random but they are statistically independent, the overall density on the data is given by [*cf.* (8.355)]

$$\mathrm{pr}(\mathbf{g}) = \int_{\mathbb{U}} d\mathbf{f}_s \int_{\mathbb{U}} d\mathbf{f}_b\; \mathrm{pr}(\mathbf{g}|\mathbf{f}_b, \mathbf{f}_s)\,\mathrm{pr}(\mathbf{f}_b)\,\mathrm{pr}(\mathbf{f}_s)\,. \tag{8.358}$$

The overall covariance matrix in this case is given by

$$\mathbf{K_g} = \overline{\mathbf{K}}_{\mathbf{n}} + \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_{\mathbf{f}_s}\boldsymbol{\mathcal{H}}^\dagger + \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_{\mathbf{f}_b}\boldsymbol{\mathcal{H}}^\dagger\,. \tag{8.359}$$

If $\mathbf{f}_s$ and $\mathbf{f}_b$ are not statistically independent, we can write $\mathrm{pr}(\mathbf{f}_b)\,\mathrm{pr}(\mathbf{f}_s) = \mathrm{pr}(\mathbf{f}_b|\mathbf{f}_s)\,\mathrm{pr}(\mathbf{f}_s)$ and do a nested average as in (8.344); the result will be that $\boldsymbol{\mathcal{K}}_{\mathbf{f}_b}$ acquires an overbar indicating that it is to be averaged over signals.