Chapter 4 An Introduction to Probability and Statistics

4.1 The Role of Probability in Inverse Problems

So far in this course, we have looked at the deterministic part of inverse problem theory, which involves examining how an “image” is transformed into the “data” via the physics of the measurement process. When this transformation is linear, the methods considered so far allow us to understand how various aspects of the image (namely the components along the appropriate singular vectors) are emphasized, de-emphasized or even obliterated within the data. When attempting to invert the data to reconstruct the image, the effect of noise and uncertainties on the measured data is to make it difficult to determine reliably the components along singular vectors with small singular values, since the naïve reconstruction method leads to a noisy quantity being divided by a small number.

The other important aspect of inverse problem theory which we have also seen is that when the data fail to tell us about some aspects of the image, it is necessary to introduce additional information in order to produce a sensible reconstruction. This can be done via **regularization methods** which attempt to select solutions which are good in some well-defined sense, such as having a small sum-squared norm, or are smooth, or are similar to what we believe the answer should be. In the framework of Tikhonov regularization, for example, we may regard solving the inverse problem as a competition between two conflicting desires, firstly the desire to minimize the residual (i.e., the misfit from the data) and secondly the desire for the solution to have a large “figure of merit”. Depending on the value of the regularization parameter, we change the relative strengths of these desires.

We now wish to consider more carefully the idea that when we reconstruct an image in an inverse problem, we have no way of being sure that the reconstruction is correct. Rather, we need a way of representing our belief that some images are more likely to be closer to the truth than others. This may either be because the data they would generate are closer to the data measured or because they more closely fit our preconceptions as to what the truth should be like. A mathematical framework which enables us to quantify the idea of the “degree of reasonable belief” is that of **subjective probability.** In this framework, probability densities represent states of knowledge over the space of possibilities and it becomes possible to formulate the general theory of inverse problems as one of statistical inference. As the data is measured, we learn more and more about the image we wish to reconstruct and at each stage, we wish to best represent our state of knowledge of the image given the available data and our preconceptions. Within this very general framework, it is possible to consider non-linear forward problems and non-additive noise processes but the size of problems which can be treated fully in this way may be rather small. Nevertheless the ideas are often very useful in the analysis of experimental data.

We begin in this chapter with a review of some of the main ideas of probability and statistics which we shall need.

4.2 Probability density functions

The **probability density function** $p_X(x)$ of a real random variable X expresses the probability that X lies in the range $a \le X < b$ in terms of the integral of the function between a and b. i.e.,

$$\Pr(a \le X < b) = \int_a^b p_X(x)\,\mathrm{d}x \tag{4.1}$$

Probability density functions are real, non-negative and normalized so that

$$\int_{-\infty}^{\infty} p_X(x)\,\mathrm{d}x = 1 \tag{4.2}$$

If we allow the probability density to be a generalized function, it is possible to use the same formalism for discrete random variables. If p_k is the probability that $X = k$ where k comes from a discrete set K, the probability density function for X is

$$p_X(x) = \sum_{k \in K} p_k \delta(x - k) \tag{4.3}$$

The generalization to several random variables is immediate. For example if X_1, X_2, ..., X_n are random variables, the probability density $p_{X_1 X_2 \ldots X_n}$ is defined so that an integral over a n-dimensional region gives the joint probability that the point $(X_1, X_2, ..., X_n)$ lies in the specified region. i.e.,

$$\begin{aligned} &\Pr(a_1 \le X_1 < b_1 \text{ and } a_2 \le X_2 < b_2 \text{ and ... and } a_n \le X_n < b_n) = \\ &\int_{a_1}^{b_1} \mathrm{d}x_1 \int_{a_2}^{b_2} \mathrm{d}x_2 ... \int_{a_n}^{b_n} \mathrm{d}x_n \, p_{X_1 X_2 \ldots X_n}(x_1, x_2, ..., x_n) \end{aligned} \tag{4.4}$$

We often use a vector notation, writing $\mathbf{X}$ for the random variable and $p_{\mathbf{X}}(\mathbf{x})$ for the probability density.

Starting from a joint probability density, we can find the probability density of a subset of the variables by integrating over all possible values of the variable(s) we do not want, e.g.,

$$p_X(x) = \int_{-\infty}^{\infty} \mathrm{d}y \int_{-\infty}^{\infty} \mathrm{d}z \, p_{XYZ}(x, y, z) \tag{4.5}$$

This process is called **marginalization** and $p_X(x)$ is called a **marginal probability density**.

Given random variables X and Y, the **conditional probability** of X given Y is defined by

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)} \tag{4.6}$$

In the joint space of possible values of X and Y, we are effectively restricting our attention to cases in which $Y = y$. Out of these, we are interested in the probability that X is equal to x.

From the definition, it is easy to see that

$$p_{XY}(x, y) = p_{X|Y}(x|y) p_Y(y) = p_{Y|X}(y|x) p_X(x) \tag{4.7}$$

This gives a relationship between the two conditional probabilities $p_{Y|X}$ and $p_{X|Y}$. As we shall see later, this is a result of fundamental importance which is called **Bayes' theorem**.

4.3 Cumulative distribution functions

The cumulative distribution function $P_X(x)$ of a single real-valued random variable X gives the probability that X is less than some specified value, i.e.,

$$\Pr(X < x_0) = P_X(x_0) \tag{4.8}$$

This is related to the probability density function $p_X(x)$ by

$$P_X(x) = \int_{-\infty}^{x} p_X(\xi) \, \mathrm{d}\xi \tag{4.9}$$

It is then easy to see that

1. $P_X(-\infty) = 0$
2. $P_X(\infty) = 1$

3. P_X is a monotonically non-decreasing function
4. $p_X(x) = \mathrm{d}P_X(x)/\mathrm{d}x$

For discrete valued random variables, the cumulative distribution function has step discontinuities. This is consistent with the delta functions in the probability density function.

In several dimensions, the cumulative distribution function $P_{X_1X_2...X_n}(x_1, x_2, ..., x_n)$ is simply the probability that $X_1 < x_1$ and $X_2 < x_2$ and ... and $X_n < x_n$. Thus

$$p_{X_1X_2...X_n}(x_1, x_2, ..., x_n) = \frac{\partial^n P_{X_1X_2...X_n}}{\partial x_1 \partial x_2 ... \partial x_n} \tag{4.10}$$

4.4 Expected values

The **expected value** of a function f of the random variable X is an average of the function values $f(x)$ weighted by the probability that X takes on the value x, i.e.,

$$\mathrm{E}[f(X)] = \int_{-\infty}^{\infty} p_X(x) f(x)\,\mathrm{d}x \tag{4.11}$$

Similarly, if we have a function of more than one random variable, the weighted average is taken over the joint probability density of the variables, i.e.,

$$\mathrm{E}[f(X_1, X_2, ..., X_n)] = \int_{-\infty}^{\infty} \mathrm{d}x_1 \int_{-\infty}^{\infty} \mathrm{d}x_2 ... \int_{-\infty}^{\infty} \mathrm{d}x_n \; p_{X_1X_2...X_n}(x_1, x_2, ..., x_n) f(x_1, x_2, ..., x_n) \tag{4.12}$$

For example, the expectation value of the product of two random variables X and Y is

$$\mathrm{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy p_{XY}(x, y)\,\mathrm{d}x\mathrm{d}y \tag{4.13}$$

The expected value of a random variable X is called the **mean** of X, and is denoted μ_X. The expected value of $(X - \mu)^2$ is called the **variance** of X and is denoted σ_X^2. The n**'th moment** m_n of X is the expected value of X^n, i.e.,

$$m_n = \mathrm{E}[X^n] = \int_{-\infty}^{\infty} x^n p_X(x)\,\mathrm{d}x \tag{4.14}$$

We see that $m_0 = 1$, $m_1 = \mu$ and $m_2 = \sigma^2 + \mu^2$.

The operation of taking the expected value is linear, hence

$$\mathrm{E}[af(X) + bg(Y)] = a\mathrm{E}[f(X)] + b\mathrm{E}[g(Y)] \tag{4.15}$$

This follows directly from the linearity of the integral.

4.5 Independent and uncorrelated random variables

Two random variables are said to be **independent** if their joint probability density is equal to the product of the individual probability densities. Thus random variables X and Y are independent if and only if

$$p_{XY}(x, y) = p_X(x) p_Y(y) \tag{4.16}$$

From the definition of conditional probability, X and Y are independent if and only if

$$p_{Y|X}(y|x) = p_Y(y) \tag{4.17}$$

Physically, this means that knowledge of the value of one of the random variables X gives no information about the value of the other random variable Y since our state of knowledge of Y conditional on knowing that $X = x$ is the same as if we had no information about X.

Similarly, a collection of random variables $X_1, X_2, ..., X_n$ is said to be independent iff their joint probability density function factorizes

$$p_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n) = p_{X_1}(x_1) p_{X_2}(x_2) ... p_{X_n}(x_n) \tag{4.18}$$

Theorem: If X and Y are independent random variables, E[XY] =E[X]E[Y]

Proof: Do as an exercise.

Two random variables X and Y are said to be **uncorrelated** if E[XY] =E[X]E[Y]. Thus independent random variables are uncorrelated, but the converse is **not** true.

Exercise: Construct two uncorrelated random variables which are not independent.

4.5.1 Example: The maximum of N independent random variables

Suppose that $X_1, ...,\ X_N$ are a set of N independent identically-distributed random variables each with probability density $p_X(x)$ and suppose that $Y = \max(X_1, ..., X_n)$. Find the probability density of Y.

It is easiest to consider the cumulative distribution function. The probability that $Y < y$ is the probability that all of $X_1,\ X_2, ..., X_N$ are less than y. Thus

$$\begin{aligned} P_Y(y) &= P_{X_1}(y) P_{X_2}(y) ... P_{X_N}(y) \\ &= \left(\int_{-\infty}^{y} p_X(x)\, \mathrm{d}x \right)^N \end{aligned} \tag{4.19}$$

Differentiating to get the probability density,

$$p_Y(y) = P_Y'(y) = N p_X(y) \left(\int_{-\infty}^{y} p_X(x)\, \mathrm{d}x \right)^{N-1} \tag{4.20}$$

4.6 Characteristic functions

The **characteristic function** of a real continuous random variable X is defined by

$$\chi_X(s) = \mathrm{E}[\exp(\mathrm{j}sX)] = \int_{-\infty}^{\infty} \exp(\mathrm{j}sx)\, p_X(x)\, \mathrm{d}x \tag{4.21}$$

This is almost the same as the Fourier transform of $p_X(x)$ except for the sign of the exponent. The inverse transform relationship is

$$p_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-\mathrm{j}sx)\, \chi_X(s)\, \mathrm{d}s. \tag{4.22}$$

If we differentiate $\chi_X(s)$ with respect to s, the effect in the integral is to multiply the integrand by jx. Thus,

$$\chi_X'(s) = \int_{-\infty}^{\infty} \mathrm{j}x\, p_X(x)\, \exp(\mathrm{j}sx)\, \mathrm{d}x \tag{4.23}$$

$$\chi_X'(0) = \int_{-\infty}^{\infty} \mathrm{j}x\, p_X(x)\, \mathrm{d}x = \mathrm{j}m_1 \tag{4.24}$$

Evaluating the derivative at $s = 0$ gives j times the mean (the first moment) of X. Successive differentiation leads to the rule

$$\chi_X^{(k)}(0) = \mathrm{j}^k m_k \tag{4.25}$$

Thus all the moments of the random variable can be derived from the characteristic function. We can also see this by expanding the exponential in the definition of the characteristic function as a power series.

$$\chi_X(s) = \mathrm{E}[\exp(\mathrm{j}sX)] = \mathrm{E}\left[\sum_{k=0}^{\infty} \frac{(\mathrm{j}sX)^k}{k!}\right] \tag{4.26}$$

$$= \sum_{k=0}^{\infty} \frac{\mathrm{j}^k \mathrm{E}\left[X^k\right]}{k!} s^k \tag{4.27}$$

$$= \sum_{k=0}^{\infty} \frac{\mathrm{j}^k m_k}{k!} s^k \tag{4.28}$$

This is the Taylor series expansion of $\chi_X(s)$ about $s = 0$. The coefficient of s^k is $\chi_X^{(k)}(0)/k!$ which again leads to the relationship (4.25).

Some important pathological cases:

- It is **not** always the case that the moments of a probability density exist and are finite. A simple example is the Cauchy probability density

  $$p_i(x) = \frac{a}{\pi\left(a^2 + x^2\right)} \tag{4.29}$$

  The second moment of this probability density is infinite.

- It is **not** the case that a complete set of moments defines the characteristic function uniquely. This is because two characteristic functions can differ by a function whose derivatives of all orders vanish at zero. Indeed it is possible to find two different probability densities which have exactly the same (finite) moments of all orders.

4.7 Probability density of the sum of independent random variables

Let X and Y be random variables with joint probability function $p_{XY}(x, y)$. We wish to find the probability density $p_Z(z)$ of the random variable Z which is the sum of X and Y.

Consider first the cumulative distribution function $P_Z(z)$. By definition,

$$P_Z(z) = \Pr(Z < z) = \Pr(X + Y < z) = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{z-x} \mathrm{d}y\, p_{XY}(x, y) \tag{4.30}$$

where the double integral is taken over the portion of the (x, y) plane for which $x + y < z$. Substituting $y' = x + y$ in the second integral yields

$$P_Z(z) = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{z} \mathrm{d}y'\, p_{XY}(x, y' - x) \tag{4.31}$$

Differentiating with respect to z yields the desired probability density function

$$p_Z(z) = \int_{-\infty}^{\infty} \mathrm{d}x\, p_{XY}(x, z - x). \tag{4.32}$$

If X and Y are **independent,** the joint density function factorizes and so

$$p_Z(z) = \int_{-\infty}^{\infty} \mathrm{d}x \, p_X(x) \, p_Y(z-x) = (p_X * p_Y)(z) \tag{4.33}$$

which we recognize as the convolution of the probability density functions. The characteristic function of Z is thus the product of the characteristic functions of X and Y

$$\chi_Z(s) = \chi_X(s)\chi_Y(s). \tag{4.34}$$

This result generalizes to the situation of more than two independent variables.

Another way of seeing that this result holds is by considering the algebra of expectation values

$$\begin{aligned}\chi_Z(s) &= \mathrm{E}[\exp(isZ)] = \mathrm{E}[\exp(is(X+Y))] \\ &= \mathrm{E}[\exp(isX)\exp(isY)] = \mathrm{E}[\exp(isX)]\,\mathrm{E}[\exp(isY)] \\ &= \chi_X(s)\chi_Y(s).\end{aligned} \tag{4.35}$$

where the factorization of the expectation value is possible because of the independence of the random variables.

Similarly if we consider $Z = aX + bY$, you should check that $\chi_Z(s) = \chi_X(as)\chi_Y(bs)$ and that the inverse transform of this characteristic function yields the probability density

$$p_Z(z) = \frac{1}{|ab|} \int p_X\left(\frac{u}{a}\right) p_Y\left(\frac{z-u}{b}\right) \mathrm{d}u \tag{4.36}$$

4.8 The Gaussian probability density

The random variable X is said to be **Gaussian** or **normally** distributed if its probability density is of the form

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4.37}$$

where μ and σ^2 are the mean and variance of X respectively. The corresponding characteristic function is

$$\chi_X(s) = \exp(\mathrm{j}s\mu)\exp\left(-\frac{1}{2}\sigma^2 s^2\right) \tag{4.38}$$

For a Gaussian random variable, the probability density is completely specified once we know the mean and the variance.

Now consider the sum of two Gaussian distributed random variables. If μ_X, μ_Y, σ_X^2 and σ_Y^2 are the means and variances of X and Y respectively, the characteristic function of the sum $Z = X + Y$ is the product of the individual characteristic functions. That is,

$$\chi_Z(s) = \exp(\mathrm{j}s\mu_X)\exp\left(-\frac{1}{2}\sigma_X^2 s^2\right) \exp(\mathrm{j}s\mu_Y)\exp\left(-\frac{1}{2}\sigma_Y^2 s^2\right) \tag{4.39}$$

$$= \exp[\mathrm{j}s(\mu_X + \mu_Y)] \exp\left[-\frac{1}{2}(\sigma_X^2 + \sigma_Y^2)s^2\right] \tag{4.40}$$

It is easy to see that Z is also Gaussian distributed. The mean of Z is $\mu_X + \mu_Y$ and the variance of Z is $\sigma_X^2 + \sigma_Y^2$. Similarly, the sum of more than two independent Gaussian distributed random variables is also Gaussian distributed. The mean of the sum is the sum of the means and the variance of the sum is the sum of the variances.

Exercise: Note that this last result is more generally true. By linearity of the expectation value, it is easy to see that the mean of the sum of two random variables is always the sum of the individual means, whether or not the random variables are independent. Show that the variance of the sum of two random variables is equal to the sum of the variances of the individual variables, provided that the random variables are **uncorrelated**.

4.9 Cumulants of a random variable

When independent random variables are added together, the mean of the sum is the sum of the means and the variance of the sum is the sum of the variances. The mean and variance are the first two of a set of quantities called **cumulants** which add together when independent random variables are added together.

The cumulants κ_n of a random variable X with probability density $P_X(x)$ are defined by

$$\log \chi_X(s) = \sum_{n=1}^{\infty} \kappa_n \frac{(\mathrm{j}s)^n}{n!} \tag{4.41}$$

They are just the coefficients of the power series expansion of the natural logarithm of the characteristic function. When two independent random variables are added together, we **multiply** together their characteristic functions. This corresponds to the **addition** of the logarithms of the characteristic functions. Thus the n'th cumulant of the sum is simply the sum of the n'th cumulants of the individual probability densities.

The first few cumulants are related to the moments as follows

$$\kappa_1 = m_1 \tag{4.42}$$

$$\kappa_2 = m_2 - m_1^2 \tag{4.43}$$

$$\kappa_3 = m_3 - 3m_2m_1 + 2m_1^3 \tag{4.44}$$

$$\kappa_4 = m_4 - 3m_2^2 - 4m_3m_1 + 12m_2m_1^2 - 6m_1^4 \tag{4.45}$$

These expressions are considerably simpler for random variables with zero means ($m_1 = 0$). To gain a physical picture of the significance of the first four moments, κ_1 is the mean, κ_2 is the variance, $\kappa_3/\kappa_2^{3/2}$ is the skewness and κ_4/κ_2^2 is the excess or kurtosis which measures whether the "skirts" of the probability density are broader ($\kappa_4 > 0$) or narrower ($\kappa_4 < 0$) than for a Gaussian of the same mean and variance.

Important note: For a Gaussian probability density, only the first two cumulants are non-zero since the logarithm of the characteristic function is a quadratic in s.

Exercise: Show that if X is a random variable and $Y = aX$ for some $a > 0$, the n'th cumulant of Y is a^n times the corresponding cumulant of X.

(*Hint:* First show that the probability density of Y is $p_Y(y) = (1/a)p_X(y/a)$.)

4.10 The central limit theorem

Let us now consider what happens when we add together N zero-mean independent identically distributed random variables. We shall assume that each of the random variables possesses n'th cumulants κ_n for every n.

Let $Z = X_1 + X_2 + ... + X_N$. It is clear that this has zero mean and that the n'th cumulant of Z is $N\kappa_n$. In particular, the variance of Z is $\sigma_Z^2 = N\kappa_2$. Consider the normalized random variable Z/σ_Z. This has unit variance and its n'th cumulant is

$$\frac{N\kappa_n}{\sigma_Z^n} = N^{1-\frac{n}{2}} \frac{\kappa_n}{\kappa_2^{n/2}} \tag{4.46}$$

As N becomes large, we see that the n'th cumulant of the normalized random variable tends to zero for all $n > 2$. We thus conclude that for large N, Z/σ_Z tends to a Gaussian random variable with zero mean and unit variance. This is a special case of the **central limit theorem**. In its more general form which applies to non-identically distributed random variables, it essentially states that

> The probability density of the sum of N well-behaved **independent** random variables tends to a Gaussian distribution whose mean is the sum of the individual means and whose variance is the sum of the individual variances

More precisely if $Z = X_1+X_2+...+X_N$, $\mu = \mu_1+\mu_2+...+\mu_N$ is the sum of the means and $\sigma^2 = \sigma_1^2+\sigma_2^2+...+\sigma_N^2$ is the sum of the variances of $X_1, X_2, ..., X_N$, then the probability density of $(Z-\mu)/\sigma$ tends to a zero-mean Gaussian distributed variable of unit variance as N becomes large.

Note that the individual probability density functions need not be Gaussian nor identically distributed.

To make this more general statement true, it is necessary to restrict the individual random variables so that each has a finite variance and that the probability for $|X_i|$ to be large is very small. These are contained in the **Lindeberg condition** which requires that for all $t > 0$,

$$\lim_{N\to\infty} \frac{1}{\sigma^2} \sum_{i=1}^{N} \int_{|x-\mu_i|>t\sigma} \mathrm{d}x\, (x-\mu_i)^2 p_i(x-\mu_i) = 0 \tag{4.47}$$

where p_i is the probability density of the i'th random variable and σ^2 is the sum of the N variances.

A rigorous proof of the central limit theorem under these general conditions is quite difficult since we do not even assume the existence of the cumulants.

Notes:

1. It is interesting to repeatedly convolve a uniform probability density with itself repeatedly to see how the probability density of the sum approaches a Gaussian. If we add together 12 independent random numbers each generated from a uniform distribution in the range $\left[-\frac{1}{2}, \frac{1}{2}\right]$, the sum closely approximates a zero-mean unit variance Gaussian distributed variable. (This is sometimes used for computer generation of normal random variables, but the method is quite slow).
2. As an example showing how the central limit theorem can fail (through violation of the Lindeberg condition), consider the sum of N identical independent Cauchy distributed random variables with

$$p_i(x) = \frac{a}{\pi\left(a^2+x^2\right)} \tag{4.48}$$

Show (as an exercise) that the variance of the Cauchy distribution is infinite and that the sum of any number of such distributions is also a Cauchy distribution and does not tend to the normal distribution.

4.11 Vector-valued random variables

A vector-valued random variable $\mathbf{X}$ with n components is simply a convenient notation for a collection of n random variables. The probability density $p_{\mathbf{X}}(\mathbf{x})$ is a joint probability density as defined above.

The **mean vector** (denoted by $\mu_{\mathbf{X}}$) is simply the vector of the mean values of the components of $\mathbf{X}$. This is the first moment of $\mathbf{X}$

$$\mu_{\mathbf{X}} = \mathrm{E}[\mathbf{X}] = \int \mathbf{x}\, p_{\mathbf{X}}(\mathbf{x})\, \mathrm{d}^n\mathbf{x} \tag{4.49}$$

The k'th component of $\mathrm{E}[\mathbf{X}]$ is $\mathrm{E}[X_k]$.

The second moment of $\mathbf{X}$ is the expectation value of products of pairs of components of $\mathbf{X}$. For an n component random vector, there are n^2 pairs which can be conveniently arranged in an $n \times n$ matrix called the correlation matrix $\Phi_{\mathbf{XX}}$

$$\Phi_{\mathbf{XX}} = \mathrm{E}[\mathbf{X}\,\mathbf{X}^t] = \int \mathbf{x}\,\mathbf{x}^t\, p_{\mathbf{X}}(\mathbf{x})\, \mathrm{d}^n\mathbf{x} \tag{4.50}$$

The kl'th component of $\Phi_{\mathbf{XX}}$ is $\mathrm{E}[X_k X_l]$. It is clear that the correlation matrix is symmetric.

Just as we defined the variance in the case of a scalar-valued random variable, we define the **covariance matrix** of a vector-valued random variable. This is also an $n \times n$ matrix $\Gamma_{\mathbf{XX}}$

$$\Gamma_{\mathbf{XX}} = E[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^t] = \Phi_{\mathbf{XX}} - \mu_{\mathbf{X}} \mu_{\mathbf{X}}^t \tag{4.51}$$

The kl'th component of $\Gamma_{\mathbf{XX}}$ is $\mathrm{E}[X_k X_l]-\mathrm{E}[X_k]\mathrm{E}[X_l]$. Like the correlation matrix, the covariance matrix is also symmetric. The diagonal elements of the covariance matrix are the variances of the random variables.

Higher order moments are more complicated to write down as the m'th moment is a rank m tensor with n^m components. The $k_1 k_2 ... k_m$'th component of the m'th moment tensor is $\mathrm{E}[X_{k_1} X_{k_2} ... X_{k_m}]$.

The multivariate form of the characteristic function is a scalar-valued function of the n dimensional vector variable $\mathbf{s}$ defined by

$$\chi_{\mathbf{X}}(\mathbf{s}) = \mathrm{E}[\exp(\mathrm{j}\mathbf{s}^t \mathbf{X})] \tag{4.52}$$

If we expand the exponential as a power series as in the scalar case, we see the successive moments appearing in the expansion. The first three terms are

$$\chi_{\mathbf{X}}(\mathbf{s}) = 1 + \mathrm{j}\mathbf{s}^t \mu_{\mathbf{X}} - \frac{1}{2!} \mathbf{s}^t \Phi_{\mathbf{XX}} \mathbf{s} - ... \tag{4.53}$$

The inverse relationship which expresses the probability density of $\mathbf{X}$ in terms of $\chi_{\mathbf{X}}(\mathbf{s})$ is

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{(2\pi)^n} \int \chi_{\mathbf{X}}(\mathbf{s}) \exp(-\mathrm{j}\mathbf{s}^t \mathbf{x}) \, \mathrm{d}^n \mathbf{s} \\ &= \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \mathrm{d}s_1 ... \int_{-\infty}^{\infty} \mathrm{d}s_n \; \chi_{\mathbf{X}}(s_1, ..., s_n) \exp[-\mathrm{j}(s_1 x_1 + s_2 x_2 + ... + s_n x_n)] \end{aligned} \tag{4.54}$$

This is essentially an n dimensional inverse Fourier transform (except for the sign of the exponent).

If the components of the vector-valued random variable $\mathbf{X}$ are **independent**, the joint probability factorizes

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1}(x_1) p_{X_2}(x_2) ... p_{X_n}(x_n) \tag{4.55}$$

As a consequence $\mathrm{E}[X_k X_l]$ $=\mathrm{E}[X_k]\mathrm{E}[X_l]$ if $k \neq l$. The covariance matrix $\Gamma_{\mathbf{XX}}$ is then diagonal with the variances on the diagonal. The characteristic function also factorizes as

$$\chi_{\mathbf{X}}(\mathbf{s}) = \chi_{X_1}(s_1) \chi_{X_2}(s_2) ... \chi_{X_n}(s_n) \tag{4.56}$$

4.12 Linear transformations and correlations

In this section we consider the generalization of the result that the characteristic function of the sum of two independent random variables is the product of the two characteristic functions. We shall see that the effect of a linear transformation is to change the **correlations** between the various components of a vector-valued random variable.

Theorem: If $\chi_{\mathbf{X}}(\mathbf{s})$ is the characteristic function of the n dimensional vector-valued random variable $\mathbf{X}$ and the m dimensional vector-valued random variable $\mathbf{Y}$ is related to $\mathbf{X}$ by the linear transformation

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \tag{4.57}$$

where $\mathbf{A}$ is an m by n matrix, the characteristic function of $\mathbf{Y}$ is given by

$$\chi_{\mathbf{Y}}(\mathbf{s}) = \chi_{\mathbf{X}}(\mathbf{A}^t \mathbf{s}) \tag{4.58}$$

Proof:

$$\begin{aligned}\chi_{\mathbf{Y}}(\mathbf{s}) &= \mathrm{E}[\exp(\mathrm{j}\mathbf{s}^t\mathbf{Y})] = \mathrm{E}[\exp(\mathrm{j}\mathbf{s}^t\mathbf{A}\mathbf{X})] = \mathrm{E}[\exp(\mathrm{j}\{\mathbf{A}^t\mathbf{s}\}^t\mathbf{X})] \\ &= \chi_{\mathbf{X}}(\mathbf{A}^t\mathbf{s}) \end{aligned} \tag{4.59}$$

It is worthwhile to consider in more detail the consequences of this deceptively simple result. We first note that it is a generalization of the result for the sum of two independent random variables in two ways. Firstly, there can be an arbitrary number of random variables contained in $\mathbf{X}$ and these need not be independent. Secondly, instead of a simple summation, we can now handle an arbitrary linear combination which can result in several random variables contained in $\mathbf{Y}$.

To see how this reduces to the previous result, suppose that $n = 2$, $m = 1$ and that $\mathbf{A} = (1\ 1)$. Then $\mathbf{Y} = \mathbf{AX}$ becomes $Y = X_1 + X_2$. By the theorem,

$$\chi_Y(s) = \chi_{\mathbf{X}}(\mathbf{A}^t s) = \chi_{\mathbf{X}}\left(\begin{pmatrix} s \\ s \end{pmatrix}\right) \tag{4.60}$$

Since the components of $\mathbf{X}$ are independent, the characteristic function factorizes, i.e.,

$$\chi_{\mathbf{X}}\left(\begin{pmatrix} s_1 \\ s_2 \end{pmatrix}\right) = \chi_{X_1}(s_1)\chi_{X_2}(s_2) \tag{4.61}$$

Hence

$$\chi_Y(s) = \chi_{X_1}(s)\chi_{X_2}(s) \tag{4.62}$$

which is just the product of the two characteristic functions. The probability densities are thus related by a convolutional relationship.

Another important consequence of this theorem may be seen by expanding the characteristic functions as power series in $\mathbf{s}$ just as in (4.53). On the left-hand side we have

$$\chi_{\mathbf{Y}}(\mathbf{s}) = 1 + \mathrm{j}\mathbf{s}^t\mu_{\mathbf{Y}} - \frac{1}{2!}\mathbf{s}^t\Phi_{\mathbf{YY}}\mathbf{s} - ... \tag{4.63}$$

and on the right-hand side,

$$\chi_{\mathbf{X}}(\mathbf{A}^t\mathbf{s}) = 1 + \mathrm{j}(\mathbf{A}^t\mathbf{s})^t\mu_{\mathbf{X}} - \frac{1}{2!}(\mathbf{A}^t\mathbf{s})^t\Phi_{\mathbf{XX}}(\mathbf{A}^t\mathbf{s}) - ... \tag{4.64}$$

$$= 1 + \mathrm{j}\mathbf{s}^t(\mathbf{A}\mu_{\mathbf{X}}) - \frac{1}{2!}\mathbf{s}^t(\mathbf{A}\Phi_{\mathbf{XX}}\mathbf{A}^t)\mathbf{s} - ... \tag{4.65}$$

Comparing these two expansions we see that

$$\mu_{\mathbf{Y}} = \mathbf{A}\mu_{\mathbf{X}} \tag{4.66}$$

$$\Phi_{\mathbf{YY}} = \mathbf{A}\Phi_{\mathbf{XX}}\mathbf{A}^t \tag{4.67}$$

These results can also be seen more directly from the definitions. For example, if $\mathbf{Y} = \mathbf{AX}$,

$$\begin{aligned}\Phi_{\mathbf{YY}} &= \mathrm{E}\left[\mathbf{YY}^t\right] = \mathrm{E}\left[\mathbf{AX}\left(\mathbf{AX}\right)^t\right] = \mathrm{E}\left[\mathbf{AXX}^t\mathbf{A}^t\right] \\ &= \mathbf{A}\mathrm{E}\left[\mathbf{XX}^t\right]\mathbf{A}^t = \mathbf{A}\Phi_{\mathbf{XX}}\mathbf{A}^t.\end{aligned}$$

Exercise: Show that the covariances are also related by

$$\Gamma_{\mathbf{YY}} = \mathbf{A}\Gamma_{\mathbf{XX}}\mathbf{A}^t \tag{4.68}$$

Thus we see precisely how a linear transformation affects the moments of the random variables. The relationship for the mean is exactly as we would expect since the process of taking an expected value is linear. Higher-order moments are similarly related via further terms in the expansions.

Exercise: Show that if $X_1, X_2, ..., X_n$ are independent random variables with means $\mu_1, \mu_2, ..., \mu_n$ and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$, then if $Z = c_1X_1 + c_2X_2 + ... + c_nX_n$, the mean of Z is $c_1\mu_1 + c_2\mu_2 + ... + c_n\mu_n$ and the variance of Z is $c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + ... + c_n^2\sigma_n^2$.

4.12.1 Physical meaning of the covariance

In order to more fully appreciate the above result, we pause to consider what the covariance matrix is telling us about the random variables that make up the vector $\mathbf{X}$. For simplicity suppose that $n = 2$ so that the pair of random variables (X_1, X_2) may be represented by a point on a plane. On successive trials, we obtain a scatter of points whose density on the plane is given by the probability density. The mean (μ_1, μ_2) is the centroid of these points. The variance of X_i is $\Gamma_{ii} = \mathrm{E}[(X_i - \mu_i)^2]$ which is (proportional to) the moment of inertia of the points when the plane is rotated about the axis $X_i = \mu_i$. These give the diagonal terms of the covariance matrix.

The off-diagonal covariance term Γ_{12} is $\mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)]$. For a given point, the product $(X_1 - \mu_1)(X_2 - \mu_2)$ is positive in the first and third quadrants (respectively negative in the second and fourth quadrants) where the two deviations $(X_1 - \mu_1)$ and $(X_2 - \mu_2)$ have the same (respectively opposite) signs. The variables are **uncorrelated** and $\Gamma_{12} = 0$ if on average the deviations are as likely to have the same signs as opposite signs. $\Gamma_{12} > 0$ and we say the variables are **positively correlated** if on average the points lie in the first and third quadrants rather than in the second and fourth quadrants. This means that if one of the variables is on one side of its mean (say $X_1 > \mu_1$), on average we expect the other variable to be on the same side of its mean (i.e., $X_2 > \mu_2$). We are not certain that this will be the case, only that as the variables become more highly positively correlated, the sign of the deviation of one of the variables becomes a more reliable indication that the sign of the other deviation is the same. The opposite holds true if $\Gamma_{12} < 0$ and the variables are **negatively correlated**. In this case, the sign of one deviation makes it likely that the other deviation is of the opposite sign.

Exercise: By expanding $\mathrm{E}\left[((X_1 - \mu_1) + \alpha(X_2 - \mu_2))^2\right]$ as a quadratic in α show that $\Gamma_{12}^2 \leq \Gamma_{11}\Gamma_{22}$ where $\Gamma_{ij} = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)]$ are the components of the covariance matrix.

(*Hint:* For all α the expectation value must be non-negative. This leads to a condition on the discriminant of the quadratic in α.)

Exercise: Show that the correlation and covariance matrices of a vector-valued random variable are positive definite matrices. (**Note:** A real-valued n by n matrix $\mathbf{A}$ is positive definite if it is symmetric and for all non-zero n dimensional column vectors $\mathbf{x}$, $\mathbf{x}^t\mathbf{A}\mathbf{x}$ is positive.)

4.13 The multivariate Gaussian and its characteristic function

First let us consider the probability density and characteristic function of n independent identically distributed Gaussian random variables with zero mean and unit variance. The probability density and characteristic function of the k'th random variable X_k are

$$p_k(x_k) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_k^2\right) \tag{4.69}$$

$$\chi_k(s_k) = \exp\left(-\frac{1}{2}s_k^2\right) \tag{4.70}$$

Since the random variables are independent, the joint probability density and characteristic function are the product of those for the individual variables

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(x_1^2 + x_2^2 + ...x_n^2)\right) \tag{4.71}$$

$$\chi_{\mathbf{X}}(\mathbf{s}) = \exp\left(-\frac{1}{2}(s_1^2 + s_2^2 + ... + s_n^2)\right) = \exp\left(-\frac{1}{2}\mathbf{s}^t\mathbf{s}\right) \tag{4.72}$$

The mean vector is $\mu_{\mathbf{X}} = \mathbf{0}$ and the correlation and covariance matrix are $\Gamma_{\mathbf{XX}} = \Phi_{\mathbf{XX}} = \mathbf{I}$ the n by n identity matrix. Now consider applying the linear transformation defined by the non-singular n by n matrix

A. i.e., we consider $\mathbf{Y} = \mathbf{AX}$. The mean and correlation matrix of $\mathbf{Y}$ are given as above by

$$\mu_{\mathbf{Y}} = \mathbf{0} \qquad \text{and} \qquad \Gamma_{\mathbf{YY}} = \Phi_{\mathbf{YY}} = \mathbf{AA}^t \tag{4.73}$$

By the theorem the characteristic function of $\mathbf{Y}$ is

$$\chi_{\mathbf{Y}}(\mathbf{s}) = \exp\left(-\frac{1}{2}(\mathbf{A}^t\mathbf{s})^t(\mathbf{A}^t\mathbf{s})\right) = \exp\left(-\frac{1}{2}\mathbf{s}^t\Gamma_{\mathbf{YY}}\mathbf{s}\right) \tag{4.74}$$

The probability density is found from the characteristic function by calculating (4.54). As shown in the appendix, the result is

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n \det(\Gamma_{\mathbf{YY}})}} \exp\left(-\frac{1}{2}\mathbf{y}^t\Gamma_{\mathbf{YY}}^{-1}\mathbf{y}\right) \tag{4.75}$$

Notice how the covariance matrix appears in the expression for the characteristic function while the inverse of the covariance matrix appears in the probability density.

The exponent of a multivariate Gaussian is a **quadratic form** in the variable $\mathbf{y}$. We may write

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n \det(\Gamma_{\mathbf{YY}})}} \exp\left(-\frac{1}{2}Q(\mathbf{y})\right) \tag{4.76}$$

where $Q(\mathbf{y}) = \mathbf{y}^t\Gamma_{\mathbf{YY}}^{-1}\mathbf{y}$. Since the matrix $\Gamma_{\mathbf{YY}}^{-1}$ is positive definite (being the inverse of a positive definite matrix), the contours $Q(\mathbf{y})$ =const form ellipsoids whose principal axes are along the eigenvectors of $\Gamma_{\mathbf{YY}}$ and whose principal axis lengths are proportional to the square roots of the eigenvalues of $\Gamma_{\mathbf{YY}}$. These contours join points of equal probability density.

If the mean of Y_k is μ_k rather than zero, the probability density and characteristic function become

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n \det(\Gamma_{\mathbf{YY}})}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu_{\mathbf{Y}})^t\Gamma_{\mathbf{YY}}^{-1}(\mathbf{y}-\mu_{\mathbf{Y}})\right) \tag{4.77}$$

$$\chi_{\mathbf{Y}}(\mathbf{s}) = \exp\left(\mathrm{j}\mathbf{s}^t\mu_{\mathbf{Y}} - \frac{1}{2}\mathbf{s}^t\Gamma_{\mathbf{YY}}\mathbf{s}\right) \tag{4.78}$$

These describe a general multivariate Gaussian random variable.

Exercise: If we start from an n dimensional multivariate Gaussian random variable $\mathbf{Y}$ and take a linear combination $\mathbf{Z} = \mathbf{AY}$, show that $\mathbf{Z}$ also has the form of a multivariate Gaussian. Thus an arbitrary linear combination of Gaussian variables is Gaussian.

4.14 Appendix: Inversion of a Gaussian characteristic function

We need to calculate the integral

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^n}\int \exp\left(-\frac{1}{2}\mathbf{s}^t\Gamma\mathbf{s} - \mathrm{j}\mathbf{s}^t\mathbf{y}\right)\,\mathrm{d}^n\mathbf{s} \tag{4.79}$$

The first step is to complete the square in the exponent. Consider the matrix analogue of a perfect square

$$\frac{1}{2}(\mathbf{s}-\mathbf{s}_0)^t\Gamma(\mathbf{s}-\mathbf{s}_0) = \frac{1}{2}\mathbf{s}^t\Gamma\mathbf{s} - \mathbf{s}^t\Gamma\mathbf{s}_0 + \frac{1}{2}\mathbf{s}_0^t\Gamma\mathbf{s}_0 \tag{4.80}$$

where we have used the fact that $\mathbf{s}^t\Gamma\mathbf{s}_0 = \mathbf{s_0}^t\Gamma\mathbf{s}$ since they are both scalars. Rearranging this gives

$$-\frac{1}{2}\mathbf{s}^t\Gamma\mathbf{s} + \mathbf{s}^t\Gamma\mathbf{s}_0 = -\frac{1}{2}(\mathbf{s}-\mathbf{s}_0)^t\Gamma(\mathbf{s}-\mathbf{s}_0) + \frac{1}{2}\mathbf{s}_0^t\Gamma\mathbf{s}_0 \tag{4.81}$$

This can be made equal to the exponent in the integrand if we set $-\mathrm{j}\mathbf{y} = \Gamma\mathbf{s}_0$ or $\mathbf{s}_0 = -\mathrm{j}\Gamma^{-1}\mathbf{y}$. The integral thus becomes

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^n}\left\{\int \exp\left(-\frac{1}{2}(\mathbf{s}-\mathbf{s}_0)^t\Gamma(\mathbf{s}-\mathbf{s}_0)\right)\,\mathrm{d}^n\mathbf{s}\right\}\exp\left(-\frac{1}{2}\mathbf{y}^t\Gamma^{-1}\mathbf{y}\right) \tag{4.82}$$

We finally consider the integral in the braces. Remembering that $\Gamma = \mathbf{A}\mathbf{A}^t$ where $\mathbf{A}$ is non-singular, we introduce the new variables

$$\mathbf{u} = \mathbf{A}^t(\mathbf{s}-\mathbf{s}_0) \tag{4.83}$$

The integral is over all of $\mathbf{s}$ space which maps to all of $\mathbf{u}$ space. The Jacobian determinant for the transformation relating the volume elements in the two spaces is

$$\mathrm{d}^n\mathbf{u} = \det(\mathbf{A}^t)\mathrm{d}^n\mathbf{s} \tag{4.84}$$

Hence

$$\begin{aligned}\int \exp\left(-\frac{1}{2}(\mathbf{s}-\mathbf{s}_0)^t\Gamma(\mathbf{s}-\mathbf{s}_0)\right)\,\mathrm{d}^n\mathbf{s} &= \int \frac{\exp\left(-\frac{1}{2}(\mathbf{u}^t\mathbf{u})\right)}{\det(\mathbf{A}^t)}\,\mathrm{d}^n\mathbf{u}\\ &= \frac{(2\pi)^{n/2}}{\det(\mathbf{A}^t)}\end{aligned} \tag{4.85}$$

Since $\det(\mathbf{A}) = \det(\mathbf{A}^t)$ and $\det(\mathbf{A})\det(\mathbf{A}^t) = \det(\Gamma)$, we see that $\det(\mathbf{A}^t) = \sqrt{\det(\Gamma)}$. Hence

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n\det(\Gamma)}}\exp\left(-\frac{1}{2}\mathbf{y}^t\Gamma^{-1}\mathbf{y}\right) \tag{4.86}$$

as claimed.

4.15 Appendix: Transformation of random variables

Suppose that X is a random variable and that $Y = f(X)$ is the random variable which is obtained by applying the function f to X. Given the probability density $p_X(x)$, we wish to determine the probability density $p_Y(y)$ of Y. It is easy to find the cumulative distribution function of Y since

$$\begin{aligned}\Pr(Y<y) &= \Pr(f(X)<y) &(4.87)\\ &= \int_{-\infty}^{\infty} u(y-f(x))\,p_X(x)\,\mathrm{d}x, &(4.88)\end{aligned}$$

where $u(x)$ is the unit step. The probability density of Y is found by differentiation

$$\begin{aligned}p_Y(y) &= \frac{\partial}{\partial y}\left(\int_{-\infty}^{\infty} u(y-f(x))\,p_X(x)\,\mathrm{d}x\right) &(4.89)\\ &= \int_{-\infty}^{\infty}\delta(y-f(x))\,p_X(x)\,\mathrm{d}x. &(4.90)\end{aligned}$$

In order to be able to apply this result, we need to be able to handle δ functions with non-trivial arguments. Recall that in distribution theory, the idea is to define the action of a distribution on a test function in such a way that the usual formal algebraic manipulations can still be carried out. Let us consider the meaning of $\delta(g(x))$ where $g(x)$ is differentiable and has a single zero at x_0 at which $g'(x_0)$ is non-zero. Given a test function $\phi(x)$, we require that

$$\langle\delta(g(x)),\phi(x)\rangle = \lim_{h\to 0}\langle\delta_h(g(x)),\phi(x)\rangle = \lim_{h\to 0}\frac{1}{h}\int_{\{x:|g(x)|<h/2\}}\phi(x)\,\mathrm{d}x \tag{4.91}$$

where $\delta_h(x)$ is equal to $1/h$ in the interval $|x| < h/2$ and is zero elsewhere. Since g has an isolated zero at x_0, for sufficiently small h, the only values of x of interest are those in a small interval around x_0. Within this interval we may approximate $g(x)$ by its Taylor series about x_0, namely

$$g(x) \approx g(x_0) + (x - x_0)\, g'(x_0) = (x - x_0)\, g'(x_0) \tag{4.92}$$

and so to this order of approximation

$$|g(x)| < \frac{h}{2} \text{ iff } |x - x_0| < \frac{h}{2\,|g'(x_0)|} \tag{4.93}$$

Thus

$$\langle \delta(g(x)), \phi(x) \rangle = \lim_{h\to 0} \frac{1}{h} \int_{|x-x_0|<\frac{h}{2|g'(x_0)|}} \phi(x)\, \mathrm{d}x = \frac{\phi(x_0)}{|g'(x_0)|}, \tag{4.94}$$

and so under these conditions,

$$\delta(g(x)) = \frac{\delta(x - x_0)}{|g'(x_0)|}. \tag{4.95}$$

If we find that $g(x)$ has several zeros $\{x_i\}$ within the interval of integration, this readily generalizes to

$$\delta(g(x)) = \sum_i \frac{\delta(x - x_i)}{|g'(x_i)|} \tag{4.96}$$

Examples

1. Suppose that the random variable Θ is uniformly distributed in the range $[0, 2\pi)$ and that $X = \tan\Theta$. Find the probability density for X and check that it is properly normalized.

   Since Θ is uniformly distributed, we see that $p_\Theta(\theta) = (2\pi)^{-1}$. By the above result,

$$p_X(x) = \int_0^{2\pi} \delta(x - \tan\theta)\; p_\Theta(\theta)\; \mathrm{d}\theta \tag{4.97}$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \delta(x - \tan\theta)\; \mathrm{d}\theta \tag{4.98}$$

   For a given value of $x > 0$, there are two values of θ within the range $[0, 2\pi)$ which satisfy $x - \tan\theta = 0$, namely $\theta_1 = \tan^{-1} x$ and $\theta_2 = \pi + \tan^{-1} x$. If we set $g(\theta) = x - \tan\theta$, we find that

$$g'(\theta) = -\sec^2\theta \tag{4.99}$$

   and so

$$|g'(\theta_1)| = |g'(\theta_2)| = \sec^2\left(\tan^{-1} x\right) = 1 + x^2 \tag{4.100}$$

   Thus

$$\delta(x - \tan\theta) = \frac{\delta\left(\theta - \tan^{-1} x\right)}{1 + x^2} + \frac{\delta\left(\theta - \pi - \tan^{-1} x\right)}{1 + x^2}. \tag{4.101}$$

   Substituting into the integral (4.98) yields

$$p_X(x) = \frac{1}{2\pi} \int_0^{2\pi} \left[\frac{\delta\left(\theta - \tan^{-1} x\right)}{1 + x^2} + \frac{\delta\left(\theta - \pi - \tan^{-1} x\right)}{1 + x^2} \right] \mathrm{d}\theta \tag{4.102}$$

$$= \frac{1}{\pi\left(1 + x^2\right)} \tag{4.103}$$

   Similarly, it is easy to check that this expression also gives the probability density for $x < 0$. The integral of $p_X(x)$ over all x yields unity, indicating that it is properly normalized.

2. Suppose that X is distributed with probability density

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{4.104}$$

and $Y = X^2$, determine the probability density of Y.

By the theorem,

$$p_Y(y) = \int \delta\left(y - x^2\right) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x \tag{4.105}$$

For $y > 0$, we see that there are two values of x, namely $\pm\sqrt{y}$ which satisfy $y - x^2 = 0$. Furthermore we find that

$$\left[\frac{\partial}{\partial x}\left(y - x^2\right)\right]_{x=\pm\sqrt{y}} = [-2x]_{x=\pm\sqrt{y}} = \mp 2\sqrt{y} \tag{4.106}$$

Hence

$$\delta\left(y - x^2\right) = \frac{\delta\left(x - \sqrt{y}\right)}{\left|-2\sqrt{y}\right|} + \frac{\delta\left(x + \sqrt{y}\right)}{\left|2\sqrt{y}\right|}. \tag{4.107}$$

Substituting into (4.105) yields

$$\begin{aligned} p_Y(y) &= \int \left(\frac{\delta\left(x - \sqrt{y}\right) + \delta\left(x + \sqrt{y}\right)}{2\sqrt{y}}\right) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \frac{1}{\sigma\sqrt{2\pi y}} \exp\left(-\frac{y}{2\sigma^2}\right). \end{aligned} \tag{4.108}$$

Exercise

Show that if $P_X(x)$ is the cumulative distribution function of a random variable X, it is possible to generate samples of X by starting with a random variable Y which is uniformly distributed within the range $[0, 1]$ and setting $X = P_X^{-1}(Y)$.

4.15.1 Extension to Multivariate Case

If we have a set of random variables $X_1, X_2, ..., X_n$ and a tranformation $f : \mathbb{R}^n \to \mathbb{R}^m$, we can find the joint probability density of $Y_1, Y_2, ..., Y_m$ where $Y_i = f_i(X_1, X_2, ..., X_n)$ by computing

$$\begin{aligned} p_{Y_1 \ldots Y_m}(y_1, ..., y_m) = \int ... \int & \delta\left(y_1 - f_1(x_1, ..., x_n)\right) ... \delta\left(y_m - f_m(x_1, ..., x_n)\right) \\ & \times p_{X_1 \ldots X_n}(x_1, ..., x_n)\, \mathrm{d}x_1 ... \mathrm{d}x_n \end{aligned} \tag{4.109}$$

An important example of the use of this theorem is to find the probability density of the sum of two random variables, i.e., when $Y = X_1 + X_2$. In this case

$$\begin{aligned} p_Y(y) &= \int\int \delta\left(y - (x_1 + x_2)\right) p_{X_1 X_2}(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \\ &= \int p_{X_1 X_2}(x_1, y - x_1)\, \mathrm{d}x_1 \end{aligned} \tag{4.110}$$

where we have carried out the integration over x_2 which collapses due to the presence of the δ function. If further we assume that the random variables are independent, so that $p_{X_1 X_2}(x_1, x_2) = p_{X_1}(x_1)\, p_{X_2}(x_2)$, this reduces to

$$p_Y(y) = \int p_{X_1}(x_1)\, p_{X_2}(y - x_1)\, \mathrm{d}x_1 \tag{4.111}$$

which is seen to be the **convolution** of the two probability densities.

4.15.2 Example: The χ^2 probability density

Consider the probability density of $Y = X_1^2 + X_2^2$ where

$$p_{X_1X_2}(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left(x_1^2 + x_2^2\right)\right) \tag{4.112}$$

By the transformation rule,

$$p_Y(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta\left(y - x_1^2 - x_2^2\right) \exp\left(-\frac{1}{2}\left(x_1^2 + x_2^2\right)\right) \mathrm{d}x_1 \mathrm{d}x_2 \tag{4.113}$$

It is convenient to change to polar coordinates. This yields

$$p_Y(y) = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} \delta\left(y - r^2\right) \exp\left(-\frac{r^2}{2}\right) r \,\mathrm{d}r \,\mathrm{d}\theta \tag{4.114}$$

Converting the delta function, we see that there the argument is zero if $r = \sqrt{y}$. At this point

$$\frac{\partial}{\partial x_1}\left(y - r^2\right) = -2r \tag{4.115}$$

Hence

$$\delta\left(y - r^2\right) = \frac{\delta\left(r - \sqrt{y}\right)}{2\sqrt{y}} \tag{4.116}$$

and so

$$\begin{aligned} p_Y(y) &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} \frac{\delta\left(r - \sqrt{y}\right)}{2\sqrt{y}} \exp\left(-\frac{r^2}{2}\right) r \,\mathrm{d}r \,\mathrm{d}\theta \\ &= \frac{1}{2} \exp\left(-\frac{y}{2}\right) \text{ for } y > 0 \end{aligned} \tag{4.117}$$

This is called the χ^2 probability density with two degrees of freedom, being the sum of squares of two independent zero-mean unit-variance Gaussian distributions. In more dimensions, the sum of squares of N independent zero-mean unit-variance Gaussian distributions has probability density

$$p_Y(y) = \frac{1}{2^{N/2}\Gamma(N/2)} y^{\frac{N}{2}-1} \exp\left(-\frac{y}{2}\right) \tag{4.118}$$

which is the χ^2 probability density with N degrees of freedom. This may readily be derived from the fact that the volume element in N dimensions for integrands with spherical symmetry may be written as

$$\mathrm{d}x_1 \mathrm{d}x_2 \ldots . \mathrm{d}x_N = \frac{2\pi^{N/2}}{\Gamma(N/2)} r^{N-1} \,\mathrm{d}r. \tag{4.119}$$

4.15.3 Characteristic function of the χ^2 probability density

Let us consider first the χ^2 density with one degree of freedom. This is the probability density of the square of a zero-mean unit-variance Gaussian distribution which is

$$p_Y(y) = \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right) \text{ for } y > 0 \tag{4.120}$$

We wish to calculate the characteristic function which is

$$\begin{aligned} \mathrm{E}\left[\exp\left(\mathrm{j}sY\right)\right] &= \int_0^\infty \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right) \exp\left(\mathrm{j}sy\right)\,\mathrm{d}y \\ &= \sqrt{\frac{1}{2\pi}} \int_{-\infty}^\infty \exp\left(\frac{1}{2}\left(2\mathrm{j}s-1\right)u^2\right)\mathrm{d}u \\ &= \sqrt{\frac{1}{2\pi}}\sqrt{\frac{2\pi}{1-2\mathrm{j}s}} = \frac{1}{\sqrt{1-2\mathrm{j}s}}. \end{aligned} \tag{4.121}$$

where we have used the change of variable $y = u^2$ and the fact that the integrand is even in the second line. For the χ^2 distribution with N degrees of freedom, we simply take the sum of N independent variables, each distributed as χ^2 with one degree of freedom. By the rule for the sum of random variables, the characteristic function is

$$\chi_Y\left(s\right) = \frac{1}{\left(1-2\mathrm{j}s\right)^{N/2}} \tag{4.122}$$

Exercises

1. Show that if $Y = aX_1 + bX_2$, then

$$p_Y\left(y\right) = \frac{1}{|ab|}\int p_{X_1X_2}\left(\frac{u}{a}, \frac{y-u}{b}\right)\mathrm{d}u. \tag{4.123}$$

   Hence find the probability density function of $3X_1 + 4X_2$ when each of X_1 and X_2 is uniformly distributed in the range $[0,1]$.

2. Find the probability density of $Z = X/Y$ if X and Y have joint probability density

$$p_{XY}\left(x,y\right) = \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right). \tag{4.124}$$

   Answer: $p_Y\left(y\right) = \frac{1}{\pi(1+y^2)}$.

3. Find the probability density of $R = \sqrt{X^2+Y^2}$ if X and Y are distributed according to the density (4.124). Answer: $p_R\left(r\right) = r\exp\left(-r^2/2\right)$, or $r > 0$.