# Chapter 4
# Multivariate Random Variables, Correlation, and Error Propagation

Of course I would not propagate error for its own sake. To do so would be not merely wicked, but diabolical.—Thomas Babington Macaulay, speech to Parliament, April 14, 1845.

## 4. Introduction

So far we have dealt with a single random variable $X$, which we can use to model collections of scalar, and similar data, such as the distance between points or the times between magnetic reversals. What we mean by "similar" is that all the data are (we assume in the model) of the same kind of thing, or measurement, so that a single, scalar r.v. $X$ can serve as the probabilistic model.

In this chapter, we generalize to this to pairs, triples, ... $m$-tuples of random variables. We may use such multiple variables either to represent vector-valued, but still similar, quantities (such as velocity, the magnetic field, or angular motions between tectonic plates); or we may use them to model situations in which we have two or more different kinds of quantities that we wish to model probabilistically. In particular, we want to have probability models that can include cases in which the data appear to depend on each other. A geophysical example of such a relationship occurs, for example, if one data type is the magnitude of an earthquake, and other is the rupture length of the fault that caused it. A compilation of such data,[1] shown in Figure 4.1, displays considerable scatter, and so is appropriate for probability modeling; but we need a model that can express the observed fact that larger earthquake magnitudes correspond to longer faults.

This plot displays a common, and important, aspect of data analysis, which is to **transform** the data in whatever way makes the relationship linear. In this case we have taken the logarithm of the rupture length—and of course, in using earthquake magnitude, the logarithm of the radiated energy (approximately).

This generalization to more than one random variable is called **multivariate probability**, though it might better be called multidimensional. We touched on the two-dimensional case in Chapters 2 and 3 as needed to discuss combinations of two independent rv's; in this chapter we extend and formalize our treatment. In particular, we describe the idea of correlation and covariance, and describe how multivariate probability is applied to the problem of propagating errors—though *not* in the sense of the quotation above.

## 4.1. Multivariate PDF's

Suppose we have an $m$-dimensional random variable $\vec{X}$, which has as components $m$ scalar random variables: $\vec{X} = X_1, X_2, \ldots, X_m$. We can easily generalize our definition of a univariate pdf to say that the r.v. $\vec{X}$ is distributed according to a **joint**

[1] Wells, D. L., and K. J. Coppersmith (1994). New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement, *Bull. Seism. Soc. Am.*, **84**, 974-1002.

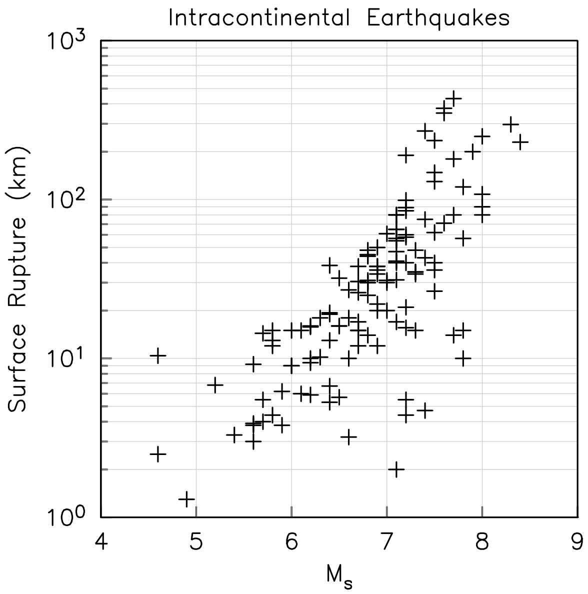

Figure 4.1

**probability density function** (which we will, as in the univariate case, just call a pdf):

$$\phi(x_1, x_2, \ldots, x_m) =_{\text{def}} \phi(\vec{x})$$

which is now the derivative of a multivariate distribution function $\Phi$:

$$\phi(\vec{x}) = \frac{\partial^m \Phi(\vec{x})}{\partial x_1 \partial x_2 \ldots \partial x_m}$$

The distribution $\Phi$ is an integral of $\phi$; if the domain of $\phi$ is not all of $m$-dimensional space, this integral needs to be done with appropriate limits. If the domain is all of $m$-dimensional space, we can write the integral as:

$$\Phi(x_1, x_2, \ldots, x_m) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_m} \phi(\vec{x})\, d^m \vec{x}$$

Given a region $R$ in $m$-dimensional space (it does not have to be any particular shape), the probability of the r.v. $\vec{X}$ falling inside $R$ is just the integral of $\phi$ over $R$

$$\mathcal{P}(X \in R) = \int_R \phi(\vec{x}) d^m \vec{x}$$

from which come the properties that $\phi$ must be everywhere nonnegative and that the integral of $\phi$ over the whole region of applicability must be 1.

Of course, it is not easy to visualize functions in $m$-dimensional space if $m$ is greater than three, or to plot them if $m$ is greater than two. Our examples will therefore focus on the case $m = 2$, for which the pdf becomes a **bivariate pdf**. Figure 4.2 shows what such a pdf might look like, plotting contours of equal values of $\phi$. (We have intentionally made this pdf a somewhat complicated one; the dots and dashed lines will be explained below). It will be evident that the probability of (say)

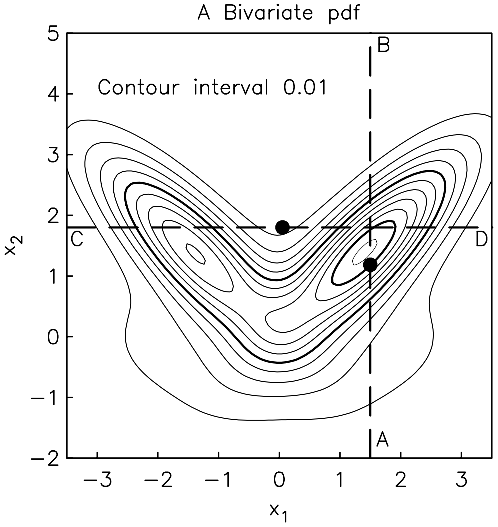

Figure 4.2

$X_2$ falling in a certain range is not unrelated to the probability of $X_1$ falling in a certain (perhaps different) range: for example, if $X_1$ is around zero, $X_2$ will tend to be; if $X_1$ is far from zero, $X_2$ will be positive. We will see how to formalize this later. It is this ability to express relationships that makes multivariate probability such a useful tool.

## 4.2. Reducing the Dimension: Conditionals and Marginals

We can, from a multivariate pdf, find two kinds of other, lower-dimensional, pdf's. We start with examples for the bivariate case, for which (obviously) the only smaller number of dimensions than $m$ is 1, so our reduction in dimension gives univariate pdf's.

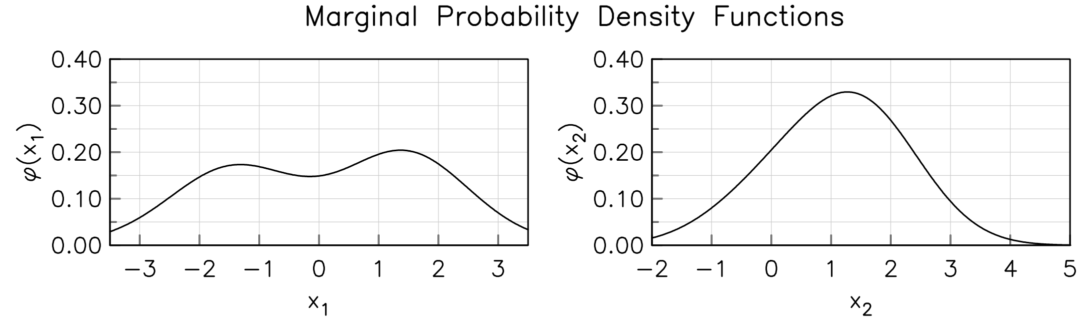

Figure 4.3

We first consider the **marginal pdf**. For either variable this is the result of integrating the bivariate pdf over the other variable. So, for example, for $X_1$ the marginal pdf is the pdf for $X_1$ irrespective of the value of $X_2$. If $\Phi(x_1, x_2)$ is the bivariate cumulative distribution, then the marginal cumulative distribution for $X_1$ is given by $\Phi(x_1, \infty)$:

$$\Phi(x_1, \infty) = \mathcal{P}(X_1 \le x_1,\, X_2 \le \infty) = \mathcal{P}(X_1 \le x)$$

But what is probably easier to visualize is the marginal density function, which comes from integrating the bivariate density function over all values of (say) $x_2$—or to put it another way, collapsing all the density onto one axis. The integral is then

$$\phi(x_1) = \int_{-\infty}^{\infty} \phi(x_1, x_2) dx_2$$

Figure 4.3 shows the marginal pdf's for the bivariate pdf plotted in Figure 4.2, with $\phi(x_{1)}$ retaining a little of the multimodal character evident in the bivariate case, though $\phi(x_2)$ does not.

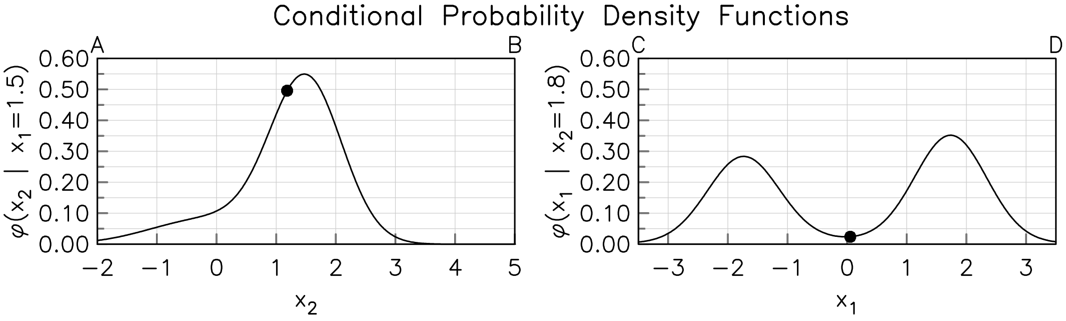

Figure 4.4

The **conditional pdf** is something quite different. It has this name because it is, for random variables, the expression of conditional probability: it gives the probability, (that is, the pdf) of (say) $X_2$ for known $x_1$; note that we write the conditioning variable in lowercase to show that it is a conventional variable, not a random one. The conditional pdf is found from $\phi(x_{1,}x_2)$ by computing

$$\phi_c(x_2) = \frac{\phi(x_1, x_2)}{\int_{-\infty}^{\infty} \phi(x_1, x_2) dx_2}$$

(We could write the conditional as $\phi_{X_2| X_1 = x_1}(x_2)$, but while this is complete it is probably confusing). We see that $x_1$ is held fixed in the integral in the denominator. The conditional pdf $\phi_c$ is essentially a slice through the multivariate pdf with one variable held fixed, normalized by its own integral so that it will integrate to 1, as it must to be a pdf. Of course we could equally well look at the conditional pdf of $X_1$ for $x_2$ held fixed—or indeed the pdf along some arbitrary slice (or even a curve) through the full bivariate pdf $\phi$. The dashed lines in Figure 4.2 show two slices for $x_1$ and $x_2$ held fixed, and Figure 4.4 shows the resulting conditional probabilities. (The dots will be explained in Section 4.5). Note that these conditional pdf's peak at much higher values than does the bivariate pdf. This illustrates the general fact that as the dimension of an r.v. increases, its pdf tends to have smaller values—as it must in order to still integrate to 1 over the whole of the relevant space.

Another name for the the marginal pdf is the **unconditional** pdf, in the sense that the marginal pdf for (say) $X_2$ describes the behavior of $X_2$ if we consider all possible values of $X_1$.

We can generalize both of these dimension-reduction strategies to more dimensions than two. If we start with a multidimensional pdf $\phi(\vec{x})$, we may either hold the variable values fixed for $k$ dimensions ($k$ being less than $m$), to get a conditional pdf of dimension $m - k$; or we may integrate $\phi$ over $k$ dimensions to get a marginal pdf of dimension $m - k$. So, for example, if we have a pdf in 3 dimensions we might:

- Integrate over one direction (it does not have to be along one of the axes) to get a bivariate marginal pdf.
- Integrate over two directions to get a univariate marginal pdf; for example, integrate over a plane, say over the $x_2$—$x_3$ plane, to get a function of $x_1$ only.
- Sample over a plane (again, it does not have to be along the axes) to get a bivariate conditional pdf.
- Sample along a line to get a univariate conditional pdf.

As we will see below, when we discuss regression, a particularly important case occurs when we take the conditional pdf for $k = m - 1$, which makes the conditional pdf univariate.

### 4.2.1. Generating Uniform Variates on the Sphere

We can apply the ideas just discussed to the problem of generating points that are uniformly distributed on the surface of a sphere, something with obvious interest in geophysics.[2] The Cartesian coordinates of points that are uniformly distributed on the surface of the unit sphere can be written as three random variables, $X_1$, $X_2$, and $X_3$. What distribution do these obey. First of all, the conditional probability distribution of $(X_1, X_2)$ for any given $X_3$ must be uniform on a circle of radius $(1 - X_3^2)^{\frac{1}{2}}$ (that is, around the circle, not within it). Next, consider the marginal distribution of any $X$, say $X_3$ (obviously, they all have to be the same). For an interval $dx_3$, the area of the corresponding slice of the sphere is proportional to $x_3$. Thus, the marginal distribution of each $X$ is uniform on $[-1, 1]$. Now suppose we generate a pair of uniform variates $U_1$ and $U_2$, each distributed uniformly between –1 and 1; then, accept any pair for which $S = U_1^2 + U_2^2 < 1$: that is, the points are inside the unit circle. Now $S$ will be uniform on [0, 1], so $1 - 2S$ is uniform on $[-1, 1]$. Hence if we set

$$\begin{array}{c} X_1 \\ X_2 \\ X_3 \end{array} = \begin{array}{c} 2U_1\sqrt{1-S} \\ 2U_2\sqrt{1-S} \\ 1-2S \end{array}$$

we see that $X_1$, $X_2$, and $X_3$ all satisfy the conditions for a uniform distribution on the sphere, provided that $1 - 2S$ is independent of $U_1/\sqrt{S}$ and $U_2/\sqrt{S}$, which it is.

## 4.3. Moments of Multivariate PDF's

We can easily generalize from moments of univariate pdf's to moments of multivariate pdf's. The zero-order moment, being the integral over the entire domain of the pdf, is still 1. But there are $m$ first moments, instead of one; these are defined by[3]

[2] Marsaglia, G. (1972). Choosing a point from the surface of a sphere, *Ann. Math. Statist.*, **43**, 645-646.

[3] *Notation alert*: we make a slight change in usage from that in Chapter 2, using the subscript to denote different moments of the same degree, rather than the degree of the moment; this degree is implicit in the number of subscripts.

$$\mu_i =_{\text{def}} \mathcal{E}[x_i] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} x_i \phi(\vec{x})\, d^m\vec{x}$$

which, as in the univariate case, expresses the location of the *i*-th variable—though not always very well.

The second moments are more varied, and more interesting, than in the univariate case: for one thing, there are $m^2$ of them. As in the univariate case, we could consider second moments about zero, or about the expected value (the first moments); in practice, nobody ever considers anything but the second kind, making the expression for the second moments

$$\mu_{ij} =_{\text{def}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}(x_i - \mu_i)(x_j - \mu_j)\phi(\vec{x})\, d^m\vec{x}$$

We can, as with univariate pdf's, describe the variance as

$$\mathcal{V}[X_i] \;=_{\text{def}}\; \mu_{ii} = \mathcal{E}[(X_i - \mu_i)(X_i - \mu_i)]$$

which as before expresses something about the spread of this variable. But the more interesting moments are the **covariances** between two variables, which are defined as

$$\begin{aligned}\mathcal{C}[X_j, X_k] \;=_{\text{def}}\; \mu_{jk} &= \mathcal{E}[(X_j - \mu_j)(X_k - \mu_k)] \\ &= \int\int\int\ldots\int(x_j - \mu_j)(x_k - \mu_k)\phi(x_1, x_2, \ldots, x_m)d^m\vec{x}\end{aligned}$$

From this, it is clear that the variances are special cases of the covariances, with the variance being the covariance of a random variable with itself:

$$\mathcal{V}[X_j] = \mathcal{C}[(X_j, X_j)].$$

Covariance has the very useful property of showing the degree of *linear* association between $X_j$ and $X_k$. To lay this out in detail:

A. If $X_j$ tends to increase linearly away from $\mu_j$ as $X_k$ increases away from $\mu_k$, then the covariance $\mathcal{C}[X_j, X_k]$ will be large and positive.

B. If $X_j$ tends to decrease linearly away from $\mu_j$ as $X_k$ increases away from $\mu_k$, then the covariance $\mathcal{C}[X_j, X_k]$ will be large and negative.

C. The covariance $\mathcal{C}[X_j, X_k]$ will be small if there is little *linear* dependence between $X_j$ and $X_k$.

For bivariate distributions we can define a particular "standardized" form of the covariance: the **correlation coefficient**, $\rho$, between $X_1$ and $X_2$

$$\rho = \frac{\mathcal{C}[X_1, X_2]}{[\mathcal{V}[X_1]\mathcal{V}[X_2]]^{\frac{1}{2}}}$$

Here "standardized" means normalized by the variances of both of the two variables. This normalization gives a quantity that varies between 1 and −1. A value of zero would mean no linear association at all; +1 or −1 would mean that $X_1$ and $X_2$ would vary exactly linearly.

## 4.4. Independence and Correlation

A special, and very important, case of a multivariate pdf occurs when the random variables $X_1, X_2, \ldots, X_m$ are **independent**; just as the probability of independent events is the product of the individual probabilities, so the pdf of independent r.v.'s can be expressed as the product of the individual pdf's of each variable:

$$\phi(\vec{x}) = \phi_1(x_1)\phi_2(x_2) \ldots \phi_m(x_m)$$

In this case the pdf of any given $X_i$ can be found independently of the distribution of all the other variables, that is to say, from $\phi_i$ alone.

If two r.v.'s $X_1$ and $X_2$ are independent then the covariance $\mathcal{C}[X_1, X_2] = 0$, and we would say that these variables are **uncorrelated**,

$$\mathcal{C}[X_1, X_2] = \int \int dx_1 dx_2 \phi_1(x_1)\phi_2(x_2)(x_1 - \mu_1)(x_2 - \mu_2)$$

$$= \int_{-\infty}^{\infty} (x_1 - \mu_1)\phi_1(x_1)dx_1 \int_{-\infty}^{\infty} (x_2 - \mu_2)\phi_2(x_2)dx_2 = 0$$

because $\mu_i = \int_{-\infty}^{\infty} x_i \phi_i(x_i)dx_i = \mathcal{E}[X_i]$ and $\int_{-\infty}^{\infty} \phi_i(x_i)dx_i = 1$.

However, the converse is not necessarily true; the covariance $\mathcal{C}[X_1, X_2]$ can be zero *without* implying statistical independence. Independence is the stronger condition on the pdf; absence of correlation refers only to a second moment of the pdf. For a slightly artificial example of no correlation but complete dependence, suppose that $X \sim N(0,1)$ and $Y = X^2$. The covariance is then

$$\mathcal{C}[X, Y] = \mathcal{E}[[X - \mathcal{E}[X]][X^2 - \mathcal{E}[X^2]]] = 0.$$

but clearly $X$ and $Y$ are not independent: $Y$ depends on $X$ exactly. What the zero covariance indicates (correctly) is that there is no linear relationship between $X$ and $Y$; indeed there is not, it is parabolic. The conditional distribution of $Y$, given $X = x$, is a discrete distribution, consisting of unit mass of probability at the point $x^2$, while the unconditional (marginal) distribution of $Y$ is $\chi^2$ with one degree of freedom.

### 4.4.1. The Multivariate Uniform Distribution

Having introduced the ideas of independence and correlation, we are in a better position to see why the generation of random numbers by computer is so difficult. The generalization to $m$ dimensions of the uniform distribution discussed in Chapter 3 would be a multivariate distribution in which:

1. Each of the $X_i$'s would have a uniform distribution between 0 and 1;
2. Any set of $n$ $X_i$'s would have a uniform distribution within an $n$-dimensional unit hypercube.
3. Each of the $X_i$'s can be regarded as independent, no matter how large $m$ becomes.

It is quite possible to satisfy (1), and not the others, as has been demonstrated by many unsatisfactory designs of random-number generators. For example, some methods fail to satisfy (2), in that for some modest value of $n$ all the combinations fall on a limited number of hyperplanes. Requirement (3) is the difficult one, both to satisfy and to test: we must ensure that no pair of $X$'s, however widely separated, have any dependence. In practice numerical generators of random numbers have some periodicity after which they begin to repeat; one aspect of designing them is to make this period much longer than the likely number of calls to the program.

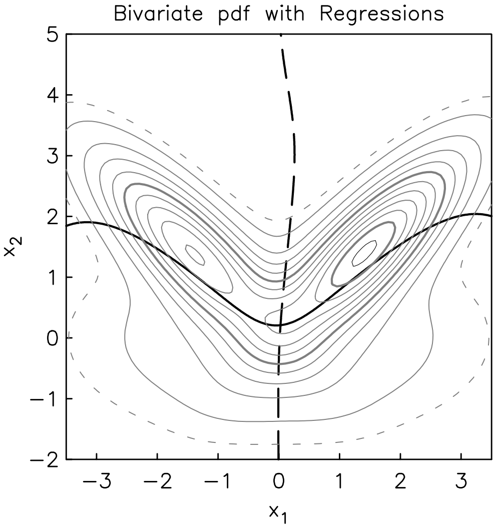

Figure 4.5

## 4.5. Regression

The idea of conditional probability for random variables allows us to introduce the concept of the regression of one variable on another, or on many others.[4]

In Figure 4.2 we showed two slices through our example bivariate pdf (the dashed lines); Figure 4.4 shows the conditional pdf's along these slices, with the expected value of $X_1$ (or $X_2$) shown by the dot in each frame of the plot. Figure 4.2 shows where these dots appear in the bivariate distribution. That the expected value of $X_1$ is so far from the peaks of the pdf is an effect of this pdf being multimodal, and with nearly similar peaks.

Now imagine finding $\mathcal{E}[X_2]$ for each value of $x_1$; that is, for each $x_1$ we get the pdf of a random variable conditional on a conventional variable; and from this, we find the expected value of that random variable; this expected value is again a conventional variable. We therefore have a function, $\mathcal{E}[X_2| x_1]$, which gives $x_2$ as a function of $x_1$. We call this the **regression of $X_2$ on $x_1$**. More usually it is called the regression of $X_2$ on $X_1$, but we prefer to make the distinction between the random and conventional variables more specific. The solid line in Figure 4.5 shows this function, which passes through one of the dots in Figure 4.2, as it should. We can see that this sort of represents the peak of the bivariate pdf. (The bivariate distribution is shown as before, with one additional contour at 0.005).

[4] Regression is another in the large array of statistical terms whose common meaning gives no clue to its technical one. The term derives from the phrase "regression to mediocrity" applied by Francis Galton to his discovery that, on average, taller parents have children shorter than themselves, and short parents taller ones. We discuss the source of this effect below.

Figure 4.5 also shows, as a dashed line, the other function we could find in this way, namely the regression of $X_1$ on $x_2$—and it is *very* different. This dashed line is, nevertheless, the correct answer to the question, "Given an $x_2$, what is the expected value of $X_1$?" You should remember that there is no unique regression: we will have a number of regression functions (usually just called regressions), depending on which variables the expected value is conditional on.

### 4.6. The Multivariate Normal Distribution

We now turn to, and spend some time exploring, the most heavily used multivariate pdf, namely the generalization to higher dimensions of the normal distribution. Our focus on this can, again, be partly justified by the Central Limit Theorem, which shows that this multivariate distribution arises from sums of random variables under general conditions: something often borne out, at least approximately, by actual data. It is also the case, as we will see, that the multivariate normal has a number of convenient properties.

The functional form of this pdf is:

$$\phi(\vec{x}) = \phi(x_1, x_2, \ldots, x_m)$$

$$= \frac{1}{(2\pi)^{m/2}|C|^{\frac{1}{2}}} \exp\left[ -\tfrac{1}{2}[(\vec{x}-\vec{\mu})\cdot C^{-1}(\vec{x}-\vec{\mu})\right]$$

We see that the mean value has been replaced by an $m$-vector of values $\vec{\mu}$. The single variance $\sigma^2$ becomes $C$, the **covariance matrix**, representing the covariances between all possible pairs of variables, and the variances of these variables themselves. $C$ is an $m \times m$ symmetric positive definite matrix, with determinant $|C| > 0$; since $C$ is symmetric, there are only $\frac{1}{2}m(m+1)$ parameters needed to define $C$: $m$ variances and $\frac{1}{2}m(m-1)$ covariances. As usual, visualization for dimensions above two are difficult, so in Figure 4.6 we show some bivariate normal distributions.

For the multivariate normal distribution we have, for the first three moments:

$$\int_{\mathcal{R}^m} \phi(\vec{x}) d^m \vec{x} = 1$$

$$\mathcal{E}[X_i] = \mu_i \qquad \mathcal{C}[X_i, X_j] = C_{ij}$$

As is the case for the univariate normal, the first and second moments completely define the pdf. This pdf also has the properties:

1. All marginal distributions are normal.
2. Given specified values of all the other variables, we get a conditional distribution that is normal.
3. If the variables are mutually uncorrelated, so that $C_{ij} = 0$ for $i \neq j$, then they are independent: in this case independence and zero correlation are equivalent.

The last result is most easily seen in the bivariate case; if $\mathcal{C}[X_{1,} X_2] = 0$ we get

$$C = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \quad \text{whence} \quad |C| = \sigma_1^2 \sigma_2^2$$

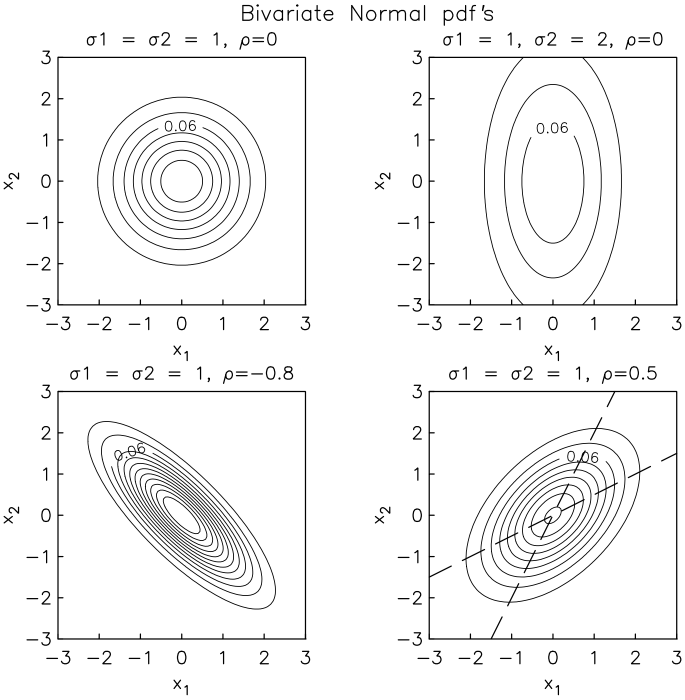

Figure 4.6

and the pdf becomes

$$\phi(x_1, x_2) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2(\sigma_1)^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_2-\mu_2)^2}{2(\sigma_2)^2}}$$

so that $X_1$ and $X_2$ are independent random variables. Contours of constant $\phi$ then form ellipses with major and minor axes that are parallel to $X_1$ and $X_2$ axes (the top two plots of Figure 4.6).

In Chapter 3 we gave a table showing, for a univariate normal, the amount of probability that we would get if we integrated the pdf between certain limits. The most useful extension of this to more than one dimension is to find the distance $R$ from the origin such that a given amount of probability (the integral over the pdf) lies within that distance. That is, for a given pdf $\phi$, we want to find $R$ such that

$$\int_0^R \phi(\vec{x}) d^m \vec{x} = p$$

To solve this for the multivariate normal, we assume that the covariance matrix is the identity (no covariances, and all variances equal to 1). Then, what we seek to do is to find $R$ such that

$$\int_0^R r^{m-1} e^{-r^2/2} dr \left[ \int_0^\infty r^{m-1} e^{-r^2/2} dr \right]^{-1} = p$$

where $r$ is the distance from the origin (in 2 and 3 dimensions, the radius). The $r^{m-1}$ term arises from our doing a (hyper)spherical integral; the definite integral is present to provide normalization without carrying around extra constants. A change of variables, $u = r^2$, makes the top integral into

$$\int_0^{R^2} u^{\frac{m-1}{2}} e^{-u/2} du$$

But remember that the $\chi^2_m$ distribution has a pdf proportional to $x^{\frac{1}{2}(m-1)}e^{-x/2}$, and we see that the integral is just proportional to the cdf of the $\chi^2_m$ pdf; if we call this $\Phi(u)$, the solution to our problem becomes $\Phi(R^2) = p$. This connection should not be too surprising when we remember that the $\chi^2_m$ distribution is just that of the sum of $m$ squares of normally-distributed rv's. Using a standard table of the cdf $\Phi$ of $\chi^2$, we can find the following values of $R$ for $p = 0.95$:

| $m$ | $R$ |
|---|---|
| 1 | 1.96 |
| 2 | 2.45 |
| 3 | 2.79 |
| 4 | 3.08 |

The first line shows the result we had before: to have 0.95 probability, the limits are (about) $2\sigma$ from the mean. For higher dimensions, the limits become farther out, since as the pdf spreads out, a larger volume is needed to contain the same amount of probability. These larger limits need to be taken into account when we are (for example) plotting the 2-dimensional equivalent of error bars.

### 4.6.1. Regression for a Multivariate Normal: Linear Regression

The regression functions for a multivariate normal take a relatively simple form, one that is widely used as a model for many regression problems. To find the regression curve, we take the conditional probability for a bivariate normal, taking $x_1$ as given; then $X_2$ has a conditional pdf:

$$\phi_{X_2|\,X_1=x_1} = \frac{1}{[2\pi\sigma_2^2(1-\rho^2)]^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\frac{[x_2-\mu_2-\rho(\sigma_1/\sigma_2)(x_1-\mu_1)]^2}{\sigma_2^2(1-\rho^2)}\right]$$

which is a somewhat messy version of the univariate pdf. From this, we see that the expected value of $X_2$ is

$$\mathcal{E}[X_2|\,X_1 = x_1] = \mu_2 + \rho(\sigma_1/\sigma_2)(x_1 - \mu_1)$$

which defines a **regression line** passing through $(\mu_1, \mu_2)$, with slope $\rho(\sigma_1/\sigma_2)$ If we now switch all the subscripts we get the regression line for $X_1$ given $X_2$, which is

$$\mathcal{E}[X_1|\,X_2 = x_2] = \mu_1 + \rho(\sigma_2/\sigma_1)(x_2 - \mu_2)$$

which passes through the same point, but has a different slope, of $\rho^{-1}(\sigma_1/\sigma_2)$. These two regression lines become the same only if the correlation coefficient $\rho$ is one. If $\rho$ is zero, the lines are at right angles. Figure 4.6 (lower right panel) shows the regression lines for the bivariate pdf contoured there.

We see that neither line falls along the major axis of the elliptical contours. This is what gives rise to the phenomenon of **regression to the mean**: if $x_1$ is above the mean, the expected value of $X_2$ given $x_1$ is less than $x_1$. So, (to use Galton's example) if parental height and offspring height are imperfectly correlated (and they are) tall parents will, on average, have children whose height is closer to the mean. But there is nothing causal about this: the same applies to tall children, who will have shorter parents on average. The same shift will be seen, and for the same reason, in repeated measurements of the same thing: remeasurement of items that initially measured below the mean will, on average, give values closer to the mean. (Assuming, again, that the variations in measurement are entirely random). When this happens, it is notoriously easy to assume that this change towards the mean shows a systematic change, rather than just the effect of random variation.

In higher dimensions, we get regression planes (or hyperplanes), but it remains true that the regression function of each variable on all the others is linear:

$$\mathcal{E}(X_i | x_1, \cdots x_{i-1}, x_{i+1}, \cdots x_m) = \sum_{j=1, j\neq i}^{n} a_j x_j + \mu_j$$

where the slopes $a_j$ depend on the covariances, including of course the variances, as in the bivariate case. So the multivariate normal pdf gives rise to linear regression, in which the relationships between variables are described by linear expressions. In fact, these linear expressions describe the pdf almost completely: they give its location and shape, with one variance needed for a complete description.

### 4.6.2. Linear Transformations of Multivariate Normal RV's

Not infrequently we form a linear combination of our original set of random variables $\vec{X}$ to produce some new set $\vec{Y}$; for example, by summing a set of random variables, perhaps weighted by different amounts. We can write any linear transformation of the original set of $m$ random variables $\vec{X}$ to produce a new set of $n$ random variables $\vec{Y}$

$$Y_j = \sum_{k=1}^{n} l_{jk} X_k$$

or in matrix notation

$$\vec{Y} = L\vec{X}$$

Note that the $l_{jk}$ are *not* random variables, nor are they moments; they are simply numbers that are elements of the matrix $L$. If the $\vec{X}$ are distributed as a joint normal pdf, the pdf for the transformed variables $\vec{Y}$ can be shown to be another joint normal with mean value

$$\vec{\nu} = L\vec{\mu} \qquad (1)$$

and covariance

$$C' = LCL^T \qquad (2)$$

We demonstrate this in the section on propagation of errors, below.

#### 4.6.2.1. Two Examples: Weighted Means, and Sum and Difference

The simplest case of such a linear combination of random variables, often used with data, is the weighted mean; from $n$ variables $X_i$ we compute

$$Y = \frac{1}{W} \sum_{i=1}^{n} w_i X_i \qquad \text{where} \qquad W = \sum_{i=1}^{n} w_i$$

We suppose that the $X_i$ are iid (independent and identically distributed), with first and second moments $\mu$ and $\sigma^2$. It is immediately obvious from (1) that $\nu = \mu$; applying (2) to the covariance matrix $C$, which in this case is a scaled version of the identity matrix, we get that the variance of $Y$ is

$$\frac{\sigma^2}{W^2} \sum_{i=1}^{n} w_i^2$$

which for all the weights the same gives the familiar result $\sigma^2/n$: the variance is reduced by a factor of $n$, or (for a normal pdf) the standard deviation by $n^{\frac{1}{2}}$.

Another simple example leads to the topic of the next section. Consider taking the sum and difference of two random variables $X_1$ and $X_2$. We can write this in matrix form as

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \tag{3}$$

The most general form for the bivariate covariance matrix $C$ is

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{4}$$

If we now apply (2) to this we obtain

$$C' = \tfrac{1}{2} \begin{pmatrix} \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2 & \sigma_1^2 - \sigma_2^2 \\ \sigma_1^2 - \sigma_2^2 & \sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2 \end{pmatrix}$$

which for the case $\sigma_1 = \sigma_2 = \sigma$ becomes

$$C' = \sigma^2 \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$$

so that the covariance has been reduced to zero, and the variances increased and decreased by an amount that depends on the correlation. What has happened is easy to picture: the change of variables (3) amounts to a 45° rotation of the axes. If $\sigma_1 = \sigma_2$ and $\rho > 0$ the pdf will be elongated along the line $X_1 = X_2$, which the rotation makes into the $Y_1$ axis. The rotation thus changes the pdf into a one appropriate to two uncorrelated variables.

There are two applications of this. One is a plotting technique called the **sum-difference plot**: as with many handy plotting tricks, this was invented by J. W. Tukey. If you have a plot of two highly correlated variables, all you can easily see is that they fall along a line. it may be instructive to replot against the sum and difference of the two, to remove this linear dependence so you can see other effects. The other application is that this kind of correlation is quite common in fitting functions to data: if the functions are very similar in shape, the fit will produce results that appear to have large variances. And so they do: but there is also a large correlation. Looking at appropriate combinations will reveal that, for example, what has a really large variance is the fit to the difference between functions; the fit to the sum will be well-determined.

### 4.6.3. Removing the Correlations from a Multivariate Normal

The example just given has shown how, for a normal pdf it is possible to find a rotation of the axes that will make the covariance matrix diagonal and the new variables uncorrelated. The question naturally arises, can this be done in general, to always produce a new set of uncorrelated variables? In this section we demonstrate that this is so.

A rotation of the axes is a special case of transforming the original variables to a new set of variables using linear combinations. For as rotation the matrix of linear combinations is square ($n = m$), and also orthogonal, with $L^T = L^{-1} =_{\text{def}} K$; the inverse matrix $K$ is also orthogonal. Then, from (1) and (2), the mean and covariance of the new variables must be $\vec{\nu}$ and

$$C' = LCK = K^T CK \tag{5}$$

Given standard results results from linear algebra it is easy to show that we can always find a new, uncorrelated, set of variables. For the transformed variables to be uncorrelated, the transformed covariance matrix has to be diagonal; from (5), we thus need to have

$$K^T CK = \text{diag}[\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2]. \tag{6}$$

Since $K$ is orthogonal, (6) is a standard spectral factorization or eigenvalue problem. For a symmetric positive definite matrix $C$ (which a covariance matrix always is), we can write the $C$ matrix as a product

$$C = U^T \Lambda U$$

where $U$ is orthogonal, $U = [\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_m]$, and $\Lambda$ is diagonal: $\Lambda = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_m]$ with all the $\lambda$'s positive; it is the case that

$$C\vec{u}_k = \lambda_k \vec{u}_k \quad \text{and} \quad \vec{u}_j \cdot \vec{u}_k = \delta_{jk}$$

so the $\vec{u}_k$ are the eigenvectors of $C$.

For our covariance problem, we let $K = U^T$, with columns given by the eigenvectors of the covariance matrix $C$; then the transformation (5) produces a diagonal matrix

$$\text{diag}\left[\lambda_1, \lambda_2, \cdots, \lambda_m\right]$$

which means that $\sigma_j^2 = \lambda_j$: the variances of the transformed set of variables are the eigenvalues of $C$. The directions $\vec{u}_j$ associated with these diagonalized $\sigma_j$ are called the **principal axes**.

While this rotation gives uncorrelated variables, there are other linear transformations that do so, for the multivariate normal at least. Another decomposition we can use is the Cholesky factorization.

$$C = LL^T$$

where $L$ is a lower triangular matrix. This yields a unit diagonal matrix if $L = K^{-1}$. For multivariate Gaussians transformation to an uncorrelated system of variables is thus not unique.

## 4.7. The Fisher Distribution

We next look at one other bivariate pdf, one that is common in some branches of geophysics. This is the **Fisher distribution**, which is the analog to the bivariate

Normal when the domain of the random variable is the surface of a sphere instead of an infinite plane. This distribution is thus useful for modeling directions on a sphere, for example directions of magnetization in paleomagnetism (for which this distribution was invented).

The pdf for this distribution is

$$\phi(\Delta, \theta) = \frac{\kappa}{4\pi \sinh \kappa} e^{-\kappa \cos \Delta}$$

To keep this expression simple, we have not included the location parameter; or rather, it is implicit in the coordinates used for the arguments of $\phi$. These are $\Delta$, which is the angular distance from the maximum of the pdf, and $\theta$, which is the azimuth—but since this does not in fact enter into the pdf, the distribution is circularly symmetric about its maximum. The location of the maximum value, with $\Delta = 0$, also corresponds to the expected value of the random variable. As with the von Mises distribution (which can be thought of as the "one-dimensional" version of the Fisher), this distribution has only a shape parameter, $\kappa$, which determines the width; this is usually called the concentration parameter. For $\kappa$ zero, the distribution is uniform over the sphere; as $\kappa$ becomes large, the pdf (near its maximum) approximates a bivariate normal with no covariance and both variances equal.

The version of the distribution just given provides the probability density over a unit angular area; angular areas on the sphere being measured in steradians, with the area of the whole (unit) sphere being $4\pi$. Another distribution derived from this is the probability density per unit of angular distance, averaging over all values of $\theta$; this is given by

$$\phi(\Delta) = \frac{\kappa}{2 \sinh \kappa} e^{-\kappa \cos \Delta} \sin \Delta$$

and, so, like the Rayleigh distribution (to which it is the analog) goes to zero at the origin. Of course, neither of these distributions incorporates possible variation in azimuth, as would be produced by unequal variances. This can be done using other distributions such as the Bingham distribution, though these have to deal with the complications of the "lobes" of the pdf wrapping around the sphere.

### 4.8. Propagation of Errors

Often we can adequately represent the actual pdf of some set of variables to be a multivariate normal; sometimes we just choose to do this implicitly by dealing only with the first two moments, the only parameters needed for a normal. Suppose all we have are first moments (means) and second moments (variances and covariances). Suppose further that we have a function $f(\vec{X})$ of the random variables; what is the multivariate normal that we should use to approximate the pdf of this function?

The usual answer to this question is done according to what is usually called the **propagation of errors**, in which we suppose not only that we can represent $\vec{X}$ adequately by its first two moments, but also suppose that we can represent the function, locally at least, by a linear approximation. All this is much simpler than what we did in Chapter 3 by finding the pdf of a rv transformed into another rv; but this simplicity is needed if we are to have manageable solutions to multivariate problems.

We start, however, by taking the function $f(\vec{X})$ to produce a single random variable. The linear approximation of the function is

$$f(\vec{X}) = f(\vec{\mu}) + \sum_{i=1}^{m} \left( \left. \frac{\partial f}{\partial x_i} \right|_{x_i = \mu_i} \right) (X_i - \mu_i) = f(\vec{\mu}) + \vec{d} \cdot (\vec{X} - \vec{\mu})^T \qquad (7)$$

where $\vec{d}$ is the $m$-vector of partial derivatives of $f$, evaluated at $\vec{\mu}$, the mean value of $\vec{X}$. The new first moment (expectation) is just the result of evaluating the function at the expected value of its argument. We can do this most easily by applying the expectation operator introduced in Chapter 2 (rather than cluttering the problem with integrals), remembering its linearity properties.

$$\mathcal{E}[f(\vec{X})] = f(\vec{\mu}) + \vec{d} \cdot \mathcal{E}[\vec{X} - \vec{\mu}]^T = f(\vec{\mu}) \qquad (8)$$

because $\mathcal{E}(\vec{X}) = \vec{\mu}$. Of course, this is only an exact result if the linear relationship (7) is a complete representation of the function: in general this is only be an approximation.

The second moment (variance) of $f(\vec{X})$ will be

$$\mathcal{V}(f(\vec{X})) = \mathcal{E}\left[ (\vec{d} \cdot (\vec{X} - \vec{\mu})^T)^2 \right] = \mathcal{E}[(\vec{d}^{\,T} \cdot (\vec{X} - \vec{\mu})) \cdot (\vec{d} \cdot (\vec{X} - \vec{\mu})^T)] =$$

$$\mathcal{E}[((\vec{X} - \vec{\mu}) \cdot \vec{d}^{\,T}) \cdot (\vec{d} \cdot (\vec{X} - \vec{\mu})^T)] = \mathcal{E}[(\vec{X} - \vec{\mu}) \cdot A \cdot (\vec{X} - \vec{\mu})^T]$$

where $A = \vec{d}^{\,T} \vec{d}$ is the $m \times m$ matrix of products of partial derivatives; in component form

$$a_{ij} = d_i d_j = \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j}$$

But we can take the expectation operator inside the matrix sum to get

$$\mathcal{V}(f(\vec{X})) = \mathcal{E}\left[ \sum_i \sum_j a_{ij} (x_i - \mu_i)(x_j - \mu_j) \right] =$$

$$\sum_i \sum_j a_{ij} \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)] = \sum_i \sum_j a_{ij} C_{ij} = \sum_{ij} d_i d_j C_{ij}$$

where $C_{ij}$ are the elements of the covariance matrix.

$$C_{ij} = \mathcal{C}(X_i - \mu_i, X_j - \mu_j)$$

We thus have an expression for the variance of the function that depends on its derivatives and on the covariance matrix.

Now consider a pair of functions $f_a(\vec{X})$, with derivatives $\vec{d}_a$ and $f_b(\vec{X})$, with derivatives $\vec{d}_b$. Then we can apply the same kind of computation to get the covariance between them as

$$\mathcal{C}[f_a(\vec{X}), f_b(\vec{X})] = \mathcal{E}[(\vec{d}_a^{\,T} \cdot (\vec{X} - \vec{\mu}) \cdot \vec{d}_b \cdot (\vec{X} - \vec{\mu})^T)]$$

$$= \vec{d}_a^T \cdot \mathcal{E}[(\vec{X} - \vec{\mu}) \cdot (\vec{X} - \vec{\mu})^T)] \cdot \vec{d}_b = \vec{d}_a^T \cdot C \cdot \vec{d}_b$$

which gives the covariance, again in terms of the derivatives of each function and the covariances of the original variables. The variances are just $\vec{d}_a^T \cdot C \cdot \vec{d}_a$ and $\vec{d}_b^T \cdot C \cdot \vec{d}_b$.

We can now generalize this to $n$ functions, rather than just two. Each function will have an $m$-vector of derivatives $\vec{d}$; we can combine these to create an $n \times m$ matrix $D$, with elements

$$D_{ij} = \frac{\partial f_i}{\partial x_j}$$

which is called the **Jacobian matrix** of the function. Following the same reasoning as before, we get that the transformation from the original covariance matrix $C$ to the new one $C'$, dimensioned $n \times n$, is

$$C' = DCD^T$$

which is of course of the same form as the transformation to a new set of axes given above in Section 4.6.2, though this is approximate, as that was not. Along the diagonal, we obtain the variances of the $n$ functions:

$$C'_{ii} = \sum_{j=1}^{m} D_{ij} \sum_{k=1}^{m} D_{ik} C_{jk}$$

### 4.8.1. An Example: Phase and Amplitude

To show how this works, we discuss the case of random variables for describing a periodic phenomenon. We can parameterize something that varies with period $T$ (not a random variable) in two ways

$$Y(t) = A \quad \cos\left(\frac{2\pi t}{T} + \theta\right) = \quad X_1 \cos\left(\frac{2\pi t}{T}\right) + X_2 \sin\left(\frac{2\pi t}{T}\right)$$

where the random variables are either the amplitude $A$ and phase $\theta$ or the cosine and sine amplitudes $X_1$ and $X_2$. The latter are sometimes called the in-phase and quadrature parts, or, if a complex representation is used, the real and imaginary parts.

Of these two representations, the first is more traditional, and in some ways easier to interpret; most notably, the amplitude $A$ does not depend on what time we take to correspond to $t = 0$. The second is preferable for discussing error (and for other purposes), because the rv's affect the result purely linearly. We may ask how the first and second moments of $X_1$ and $X_2$ map into the same moments of $A$ and $\theta$. The relationship between these is

$$A = f_1(X_1, X_2^2) = (X_1^2 + X_2^2)^{\frac{1}{2}} \qquad \theta = f_2(X_1, X_2^2) = \arctan\left(\frac{X_1}{X_2}\right) \tag{9}$$

from which we can find the components of the Jacobian matrix to be

$$d_{11} = \frac{\partial f_1}{\partial x_1} = \frac{x_1}{a} \qquad d_{12} = \frac{x_2}{a}$$

$$d_{21} = \frac{\partial f_2}{\partial x_1} = \frac{x_2}{a^2} \qquad d_{22} = \frac{x_1}{a^2}$$

where we have used $a$ for the amplitude to indicate that this is not a random variable but a conventional one, to avoid taking a derivative with respect to a random variable.

If we assume that the variances for the $X$'s are as in (4), we find that the variances for $A$ and $\theta$ are

$$\sigma_A^2 = \frac{x_1}{a}\sigma_1^2 + \frac{x_2}{a}\sigma_2^2 - \frac{x_1x_2}{a^2}\rho\sigma_1\sigma_2 \qquad \sigma_\theta^2 = \frac{x_2^2}{a^4}\sigma_1^2 + \frac{x_1^2}{a^4}\sigma_2^2 - \frac{x_1x_2}{a^4}\rho\sigma_1\sigma_2$$

which, if $\sigma_1 = \sigma_2$ and $\rho = 0$, reduces to

$$\sigma_A = \sigma \qquad \sigma_\theta = \frac{\sigma}{a}$$

which makes sense if thought about geometrically: provided $a$ is much larger than $\sigma_1$ and $\sigma_2$, the error in amplitude looks like the error in the cosine and sine parts, while the error in phase is just the angle subtended. However, if the square root of the variance is comparable to the amplitude, these linear approximations fail. For one thing, the result (8) no longer holds: the expected value of the amplitude is not $(X_1^2 + X_2^2)^{\frac{1}{2}}$, but is systematically larger; in fact for the $X$'s both zero, $A$ has the Rayleigh distribution. And the pdf for the angle $\theta$, also, cannot be approximated by a normal distribution. This is a special case of a more general result: if you have a choice between variable with normal errors and another set which is related to them nonlinearly, avoid the nonlinear set unless you are prepared to do the transformation of the pdf properly, or have made sure that the linear approximation will be valid.

## 4.9. Regression and Curve Fitting

As a transition to the next chapter, on estimation, we return to the problem of finding the regression curve. We compare this with a superficially similar procedure that is often (confusingly) also called regression, but is in fact conceptually different, although mathematically similar (hence the use of the name).

In regression as we defined it above, we assume a pdf for at least two variables, and ask for the expected value of one of them when all the others are fixed. Which one we do not fix is a matter of choice; to go back to Figure 4.1, it would depend on whether we were trying to predict magnitude from rupture length, or the other way around.

The other usage of the term regression is for fitting some kind of function to data. We suppose we have only one thing that should be represented by a random variable; and that this thing depends on one or more conventional variables through some functional dependence that we wish to determine. The regression curve (or more generally surface) is then the function we wish to find. But there is not the symmetry we have in the regression problem; only one variable is random.

Most elementary discussions of the analysis of experimental data assume a model that amounts to this, but one which uses an approach we described in Chapter 1, and chose not to use: that is to say that we should regard data $y$ as being the sum of a function $f_p(\vec{x})$ and a random variable $\vec{E}$. Here $f_p$ is a function that depends on some parameters $p$ and takes as argument some variables $\vec{x}$, and $\vec{E}$ is the "error".

In our terminology, $\vec{x}$ is a vector of conventional variables and the function $f_p$ produces a conventional variable; the error is modeled as a random variable. A simple example would be if we measure the position of a falling object (with zero initial velocity) as a function of known times $\vec{t}$. We would write this as

$$\vec{Y} = \tfrac{1}{2} g\vec{t}^{\,2} + \vec{E} \tag{10}$$

where we take the $t^2$ to represent the squares of the individual $t$'s. The parameter of the function is $g$, the gravitational acceleration. In our framework, $\vec{E}$ is a random variable, and so is $\vec{Y}$. Now assume the pdf of $\vec{E}$ to be $\phi(y)\vec{I}$, where $I$ is the identity matrix, so the individual $Y$'s are independent and identically distributed. Then (10) can be written as a distribution for $Y$:

$$Y_i \sim \phi(y - \tfrac{1}{2} g t_i^2) \tag{11}$$

so that $g$, and the known values $t_i$, appear as parameters in a pdf—which, as conventional variables, is where they must appear. So the problem of finding $g$, given the values for $\vec{t}$ and $\vec{Y}$, becomes one of estimating parameters in a pdf, something which we will examine extensively in the next chapter.

You should by now be able to see that this is quite a different setup from the one that led us to regression. In particular, since $t$ is not a random variable, there is no way in which we can state a bivariate pdf for $t$ and $Y$. We could of course contour the pdf for $Y$ against $t$, but these contours would not be those of a bivariate pdf; in particular, there would be no requirement that the integral over $Y$ and $t$ together be equal to 1. However, it is also easy to see (we hope) that we can engage in the activity of finding a curve that represents $\mathcal{E}(Y)$ as a function of $t$ and the parameter $g$; and that this curve, being the expected value of $Y$ given $t$, is found in just the same way as the regression curve for one random variable conditional upon the other one. So in that sense "regression" is equivalent to "curve fitting when only one variable is random [e.g., has errors)". But it is important to remember that the similar solutions of these two problems come from a quite different underlying structure.