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 the House of Commons, April 14, 1845

4.1 Introduction

So far we have mostly discussed the theory needed to model sets of similar data, such as the distance between points or the times between magnetic reversals; we use "similar" to indicate that the data are all the same kind of thing. Then a single, scalar, rv X is adequate for a probabilistic model.

We now generalize this to pairs, triples, ... m-tuples of random variables. We may use such multiple variables either to represent vector-valued, but still similar, quantities (for example, velocity); or we may use them to model data for different kinds of quantities. In particular, we introduce probability models that can model apparent dependencies between different data. Figure 4.1 shows an example, in which the magnitude of an earthquake is related to the rupture length of the fault that caused it. These data are scattered, so a probability model is appropriate; but we want our model to express the observed fact that larger earthquake magnitudes correspond to longer faults.

This plot displays an important aspect of analyzing paired data, namely the value of **transforming** the data in some way to make any relationship into a linear one: in this case by taking the logarithm of the rupture length.

Generalizing to more than one random variable brings us to **multivariate probability**, though it might better be called multidimensional. In or-

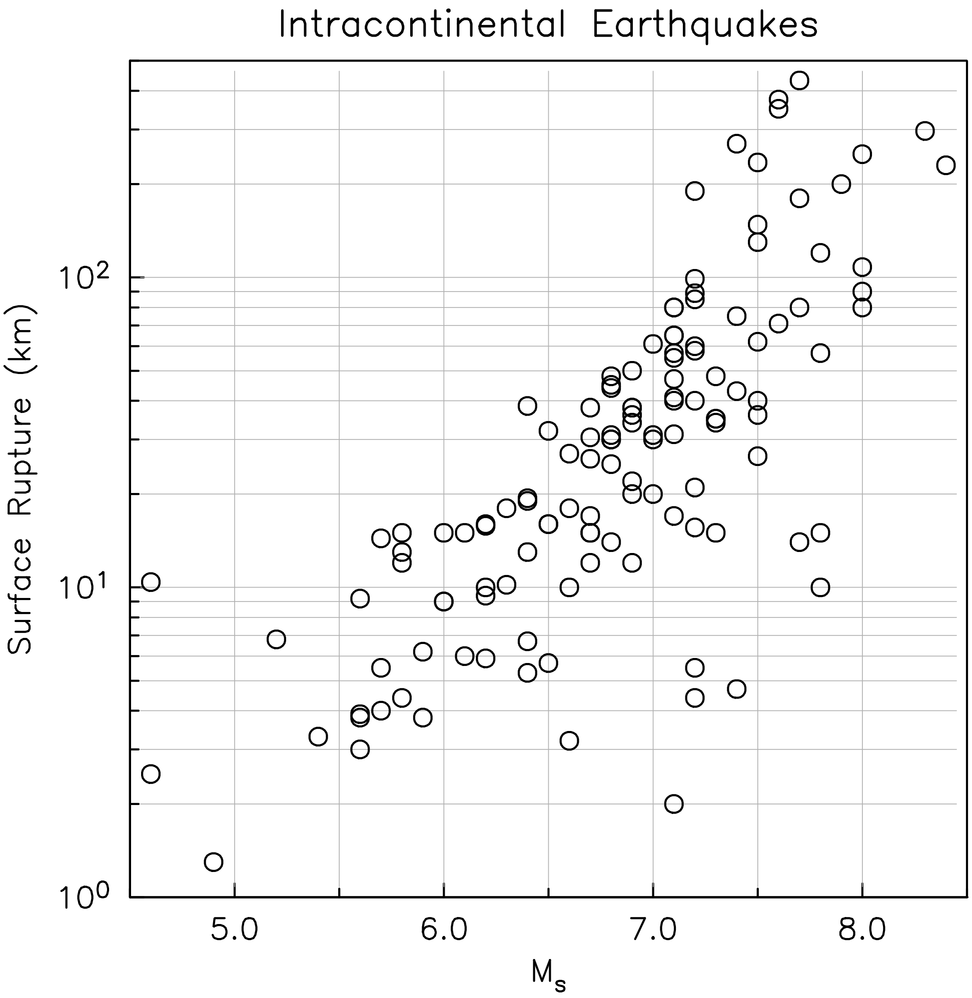

Figure 4.1: Length of surface rupture and magnitude for 128 earthquakes in continental regions, from the compilation by Triep and Sykes; most of the rupture-length data are from *Wells and Coppersmith* [1994].

der to discuss combinations of two independent rv's we needed to introduce the two-dimensional case in Section 2.11; in this chapter we extend and formalize our treatment. In particular, we describe correlation and covariance, and also how to apply multivariate probability to propagating errors – though *not* in the sense of the chapter epigraph.

4.2 Multivariate PDF's

We introduce an m-dimensional (vector) 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 the definition of a univariate pdf to a **joint probability density function** (which we also call a pdf) that describes the distribution of $\vec{X}$. This pdf is

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

which is the derivative of a **multivariate distribution function** Φ:

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

The distribution Φ is an integral of ϕ; if the domain of ϕ 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:

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

If we have a region R of any shape in m-dimensional space, the probability of the rv $\vec{X}$ falling inside R is just the integral of ϕ over R

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

This implies that ϕ must be nonnegative everywhere and that the integral of ϕ over the whole region of applicability must be one.

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 ϕ. It will be evident that the probability of (say) 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 also tend to be; if X_1 is far from zero, X_2 will be positive. This ability to express relationships is what makes multivariate probability so useful.

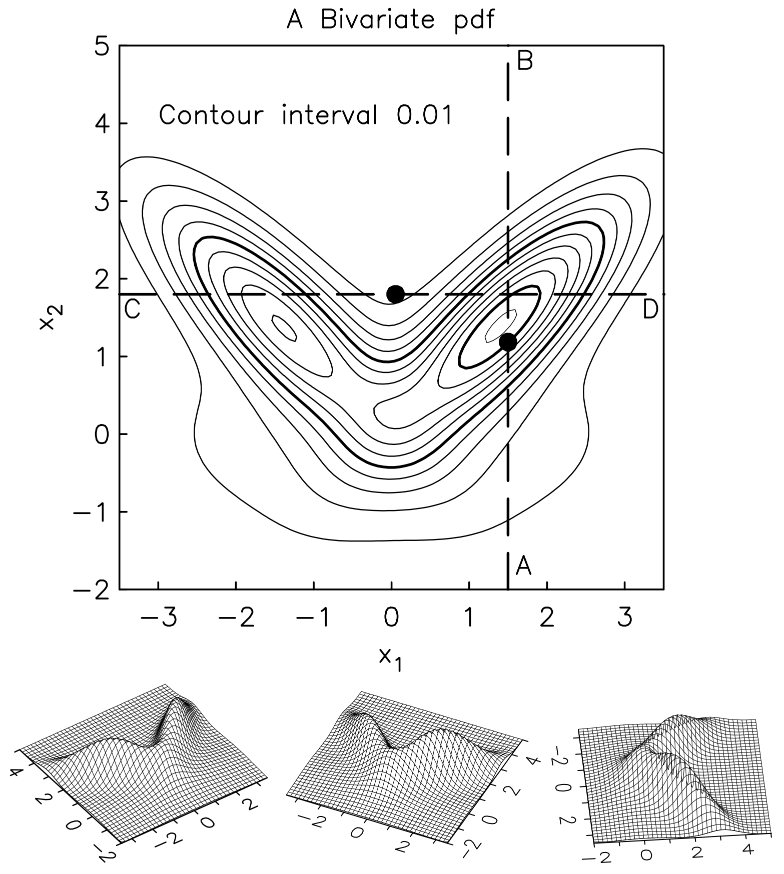

Figure 4.2: Example of a somewhat complicated bivariate pdf, shown using contours, The dashed lines and dots are explained in Sections 4.3 and 4.6,

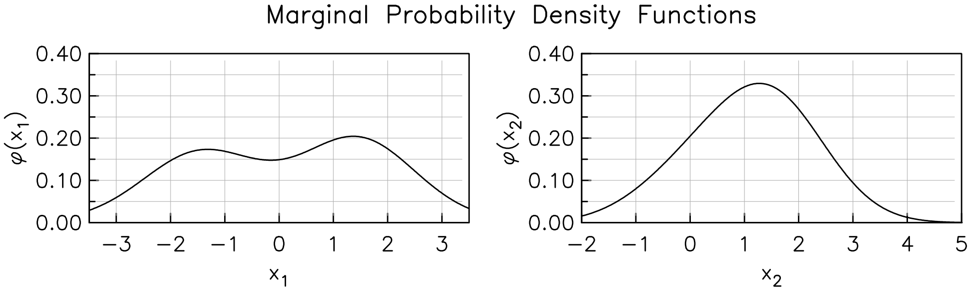

Figure 4.3: Marginal pdf's for the bivariate pdf of Figure 4.2.

4.3 Reducing the Dimension: Conditionals and Marginals

From a multivariate pdf we can find two kinds of lower-dimensional pdf's: Since our examples are for the bivariate case, the only smaller number of dimensions is one, to univariate pdf's.

The first reduced-dimension pdf is the **marginal pdf**; in the bivariate case this is the pdf for one variable irrespective of the value of the other. If $\Phi(x_1,x_2)$ is the bivariate cumulative distribution, then the **marginal cumulative distribution function** for X_1 would be given by $\Phi(x_1,\infty)$:

$$\Phi(x_1,\infty) = \Pr(X_1 \le x_1, X_2 \le \infty) = \Pr(X_1 \le x)$$

It is easier to visualize 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. So for a bivariate pdf, the two marginal pdf's are

$$\phi(x_1) = \int_{-\infty}^{\infty} \phi(x_1,x_2)\,dx_2 \quad \text{and} \quad \phi(x_2) = \int_{-\infty}^{\infty} \phi(x_1,x_2)\,dx_1$$

Figure 4.3 shows the marginal pdf's for the bivariate pdf plotted in Figure 4.2; while $\phi(x_1)$ retains a little of the multimodal character evident in the bivariate pdf, $\phi(x_2)$ does not.

The other lower-dimension pdf is the **conditional probability density function** – which is very different from the marginal. The conditional pdf is so called because it expresses conditional probabilities, something we did for events in Section 2.4 but could not, up to now, apply to random variables. It gives the probability density for X_2 (say) given a known x_1.

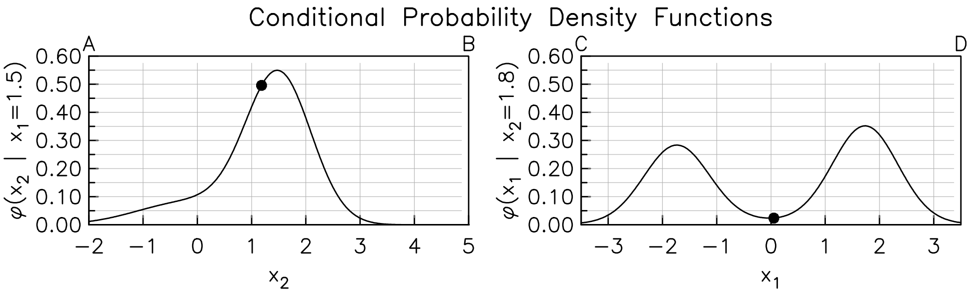

Figure 4.4: Examples of conditional pdf's for the bivariate pdf of Figure 4.2. The dots are explained in Section 4.6.

Note that we write x_1, which is called 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

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

Strictly speaking we might want to write the conditional pdf as $\phi_{X_2|X_1=x_1}(x_2)$, but while this is more complete it is probably also more confusing. Note that x_1 is held fixed in the integral in the denominator. The conditional pdf ϕ_c is essentially a slice through the multivariate pdf, holding one variable fixed, and normalizing by its own integral to make the integral of ϕ_c equal to one, as must be true for a pdf. There is nothing special about making x_1 the conditioning variable: we could equally well define the conditional pdf of X_1 for x_2 held fixed – indeed we could define a conditional of the pdf by taking a slice of the bivariate pdf in any direction, or indeed along a curve. In Figure 4.2 the dashed lines show two slices for x_1 and x_2 fixed; Figure 4.4 plots the resulting conditional probabilities. 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 rv increases, its pdf tends to have smaller values – as it must in order to still integrate to one over the whole of the relevant space.

The **unconditional** pdf is another name for the marginal pdf, since the marginal pdf for (say) X_2 describes the behavior of X_2 if we consider all possible values of X_1.

We can easily generalize these dimension-reduction procedures 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 ϕ over k dimensions to get a marginal pdf of dimension $m-k$. So, for example, if we have a pdf in three dimensions we might:

1. Integrate over one direction (it does not have to be along one of the axes) to get a bivariate marginal pdf.

2. 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.

3. Take a slice over a plane (again, it does not have to be along the axes) to get a bivariate conditional pdf.

4. Take a slice along a line to get a univariate conditional pdf (say, what is the pdf of X_3 for specified values of x_1 and x_2).

When we discuss regression in Section 4.6, we will see that the last example, in which the conditional pdf is univariate, ($k = m-1$) is especially important.

4.3.1 Uniformly-Distributed Variates on the Sphere

We can apply the ideas of the previous section to a problem of geophysical interest, namely generating points that have an equal probability of being anywhere on the surface of a sphere: anther type of uniform distribution [*Marsaglia*, 1972]. We start by noting that the Cartesian coordinates of such points are a vector of three random variables, X_1, X_2, and X_3. Then we can ask, what distribution do these obey for a uniform distribution on a unit sphere?

For such a distribution the conditional pdf of (X_1,X_2) for any given x_3 must be uniform on the circumference of a circle of radius $r = (1-x_3^2)^{1/2}$. Next, consider the marginal distribution of any particular X, say X_3 (obviously, they all have to be the same): that is, what is the marginal pdf along the x_3 axis if we collapse the entire sphere onto it? A constant pdf on the surface of the sphere gives a marginal pdf that is proportional to the area falling in an interval dx_3: this is just $2\pi r$, with r given as above. The area of the corresponding slice of the sphere is proportional to x_3. Therefore, 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; and then accept any pair for which $S = U_1^2 + U_2^2 \leq 1$: 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

$$X_1 = 2U_1\sqrt{1-S} \qquad X_2 = 2U_2\sqrt{1-S}$$

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.

4.4 Moments of Multivariate PDF's

We can easily generalize 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 one. But there will be m first moments, instead of one; these are defined by[1]

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

which, as in the univariate case, more or less expresses the likely location of the i-th variable.

The second moments become more varied, and more interesting, than for 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

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

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

$$\mathscr{V}[X_i] \overset{\text{def}}{=} \mu_{ii} = \mathscr{E}[(X_i - \mu_i)(X_i - \mu_i)]$$

and as before the variance roughly indicates the spread of the variable it applied to.

[1] *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.

But the new, and more interesting, moments are the **covariances** between two variables, defined as

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

The term "covariance" comes from the fact that these values express how much the rv's "co-vary"; that is, vary together. 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:

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

The covariance expresses the degree of *linear* association between X_j and X_k; in more detail:

1. If X_j tends to increase linearly away from μ_j as X_k increases away from μ_k, then the covariance $\mathscr{C}[X_j, X_k]$ will be large and positive.

2. If X_j tends to decrease linearly away from μ_j as X_k increases away from μ_k, then the covariance $\mathscr{C}[X_j, X_k]$ will be large and negative.

3. If there is little *linear* dependence between X_j and X_k the covariance $\mathscr{C}[X_j, X_k]$ will be small.

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

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

Here "standardized" means normalized by the variances of both of the two variables, which means that $-1 \le \rho \le 1$. If $\rho = 0$, there is no linear association at all; for $\rho = \pm 1$ X_1 and X_2 vary together exactly linearly, with no scatter.

4.5 Independence and Correlation

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

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

so that for independent rv's, the pdf of each one can be give independently of the distribution of any of the others.

If two rv's X_1 and X_2 are independent then the covariance $\mathscr{C}[X_1,X_2] = 0$, and these variables are termed **uncorrelated**:

$$\left.\begin{aligned} \mathscr{C}[X_1,X_2] &= \int\int dx_1 dx_2 \phi_1(x_1)\phi_2(x_2)(x_1-\mu_1)(x_2-\mu_2) \\ &= \int_{-\infty}^{\infty}(x_1-\mu_1)\phi_1(x_1)\,dx_1 \int_{-\infty}^{\infty}(x_2-\mu_2)\phi_2(x_2)\,dx_2 \end{aligned}\right. = 0$$

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

However, the converse is not necessarily true; the covariance $\mathscr{C}[X_1,X_2]$ can be zero *without* implying statistical independence. Independence is the stronger condition, since correlation refers only to second moments 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

$$\mathscr{C}[X,Y] = \mathscr{E}\left[[[X-\mathscr{E}[X]][X^2-\mathscr{E}[X^2]]\right] = 0$$

but clearly X and Y are not independent: Y depends on X exactly. What the zero covariance indicates (correctly) is that there should be no linear relationship between X and Y; and in this case that is indeed so: the dependence is parabolic. Also for this case 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 χ_1^2 with one degree of freedom.

4.5.1 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 Section 3.2 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.

The many unsatisfactory designs of random-number generators demonstrate that it is quite possible to satisfy the first condition without satisfying the others. 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 most 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 part of designing them is to make this period much longer than the likely number of calls to the program.

4.6 Regression

The idea of conditional probability for random variables allows us to introduce regression, which relates the expected value of some variables to the values specified for others.[2] Figure 4.2 shows, as dashed lines, two slices through our example of a bivariate pdf, and Figure 4.4 shows the conditional pdf's taken along these slices. From these conditional pdf we can find the expected value of X_1 (or X_2), indicated in the figure by the large dot; since this corresponds to a particular value of both variables, we can also show the dots in Figure 4.2 to plot where they fall relative to the bivariate distribution. Note that the expected value of X_1 is far from the peaks of the pdf, but this is because the pdf is multimodal.

Now imagine finding $\mathscr{E}[X_2]$ for each value of x_1; that is, for each x_1 we get the conditional pdf for X_2; and from this pdf, we find the expected value of X_2; but $\mathscr{E}[X_2]$ is again a conventional variable. We have thus created a function, $\mathscr{E}[X_2|x_1]$, which gives x_2 as a function of x_1. We call this

[2] Regression is yet another statistical term whose common meaning gives no clue to its technical one. The terminology ultimately 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 – an effect we discuss in Section 4.7.1, using there the other name, "regression to the mean".

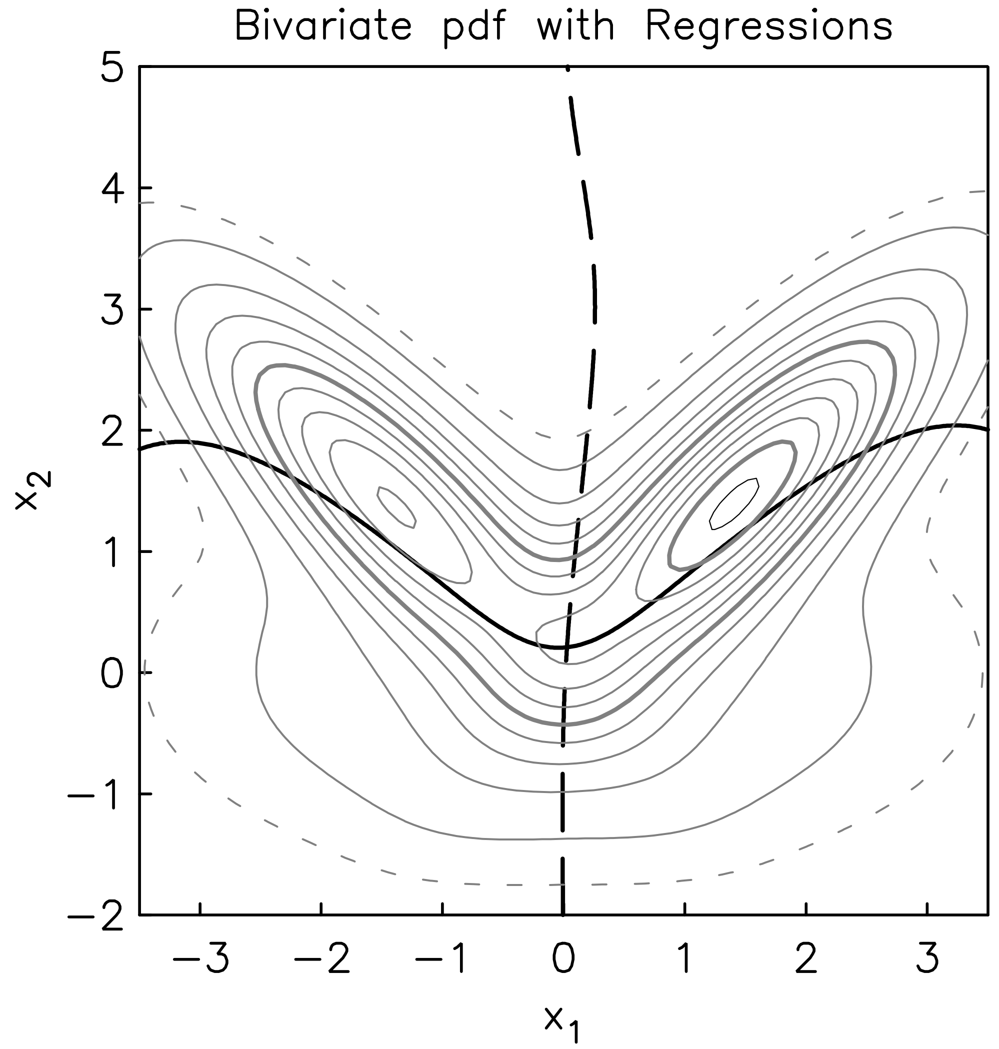

Figure 4.5: The gray lines show the bivariate distribution of Figure 4.2, with one additional contour at 0.005. The black curves show the regression of X_2 on x_1 (solid) and X_1 on x_2 (dashed): they are *very* different.

function the **regression of X_2 on x_1**.[3] The heavy line in Figure 4.5 shows this function, which, as it should, passes through one of the dots in Figure 4.2. This function follows the highest values of the bivariate pdf, at least approximately.

Figure 4.5 also shows the other function we could find in this way, namely the regression of X_1 on x_2; that is, $\mathscr{E}(X_1|x_2)$. Although this function is very different, it does give the correct answer to the question, "Given a specified x_2, what is the expected value of X_1?" Here, and in general, there is not any single regression: we can have many regression functions (usually just called regressions) each depending on what variables the expected value is conditional on. What variables we want to specify, and which ones we then want to find the expected value of, is another example of something that has to be decided for the particular problem being investigated. A seismological example [*Agnew*, 2010] is relating two magnitude scales after a change in how magnitude was defined: we can ask how to find the expected value of a new-style magnitude given an old-style value, or the other way around.

4.7 The Multivariate Normal Distribution

In this section we examine, in some detail, the most heavily used multivariate pdf, which generalizes the normal distribution to higher dimensions. Our focus on this particular pdf can be partly justified by the Central Limit Theorem, which shows that in general sums of random variables give this multivariate pdf. Also, the multivariate normal has several convenient properties.

The functional form of this pdf is:

$$\phi(\vec{x}) = \phi(x_1, x_2, \ldots, x_m) = \frac{1}{(2\pi)^{m/2}|C|^{1/2}} \exp\left[-\tfrac{1}{2}[(\vec{x}-\vec{\mu})\cdot C^{-1}(\vec{x}-\vec{\mu})\right]$$

We see that the mean value has been replaced by an m-vector of values $\vec{\mu}$. The single variance σ^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, $|C| > 0$; because C is symmetric, we need $\tfrac{1}{2}m(m+1)$ parameters (m variances and $\tfrac{1}{2}m(m-1)$ covariances) to define it

[3] 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.

completely. As usual, this pdf is difficult to visualize for dimensions above two; Figure 4.6 shows some examples of this pdf for two dimensions (the bivariate Normal).

The first three moments of the multivariate normal distribution are

$$\int_{\Re^m} \phi(\vec{x}) d^m \vec{x} = 1 \qquad \mathscr{E}[X_i] = \mu_i \qquad \mathscr{C}[X_i, X_j] = C_{ij}$$

As for the univariate normal, the first and second moments completely define the pdf. This pdf (in m dimensions) also has the following properties:

1. All marginal distributions are normal.

2. All conditional distributions, of whatever dimension, are normal.

3. If the variables are mutually uncorrelated, so that $C_{ij} = 0$ for $i \neq j$, then they are also independent: for the multivariate normal, independence is equivalent to zero correlation.

The last result is easily demonstrated for the bivariate case; if $\mathscr{C}[X_1, X_2] = 0$

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

and the pdf becomes

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

and if $\phi(x_1, x_2) = \phi(x_1)\phi(x_2)$ the X_1 and X_2 are independent random variables, and, as in the top two plots of Figure 4.6, the contours of constant ϕ will be ellipses with major and minor axes parallel to the x_1 and x_2 axes.

Table 3.3 showed, 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 ϕ, we want to find R such that

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

To get a single number for R, we need to assume that the covariance array is the identity matrix (no covariances, and all variances equal to one).

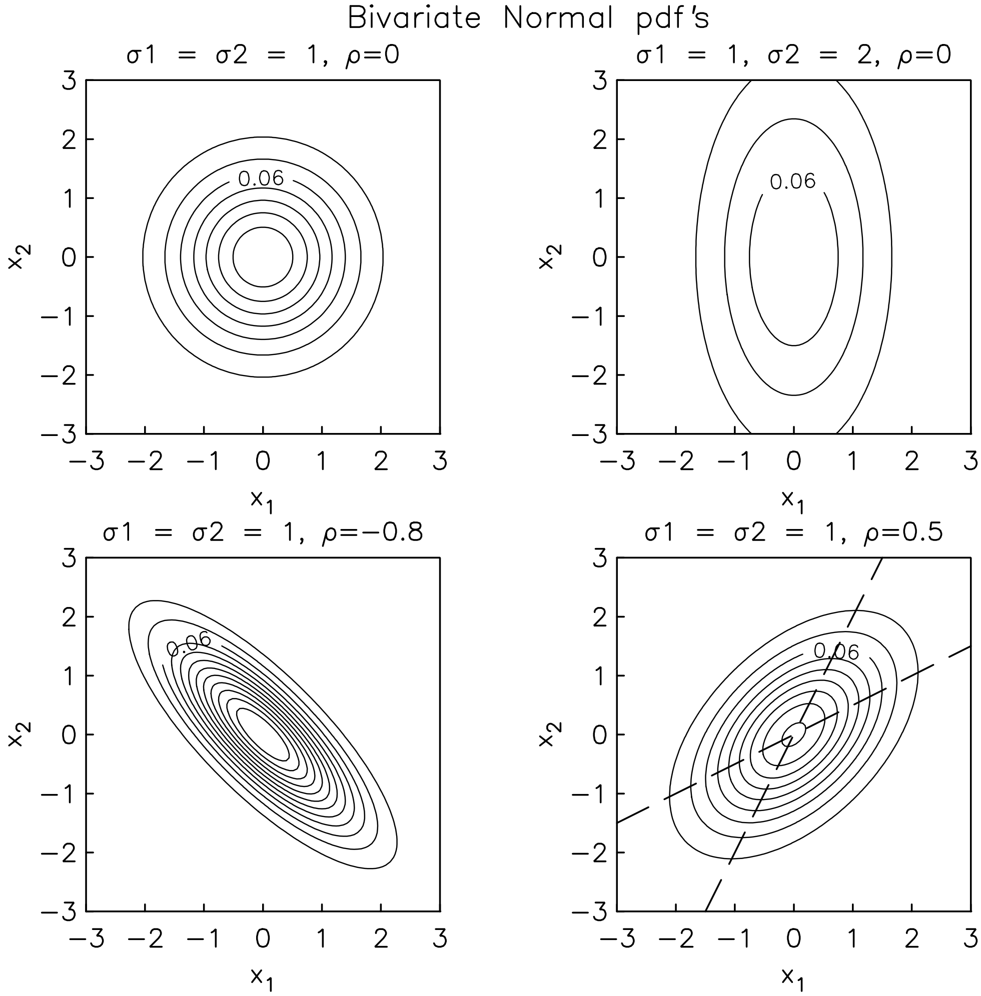

Figure 4.6: Contour plots of bivariate Normal distributions for various combinations of the second moments; the first moment is always zero. The dashed lines in the lower right panel are the two regression lines for this pdf.

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. 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 numerator into

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

Now remember that the χ^2_m distribution has a pdf proportional to $x^{\frac{1}{2}(m-1)}e^{-x/2}$, so we see that the integral is just proportional to the cdf of the χ^2_m pdf; if we call this $\Phi(u)$, we can write the equation we want to solve as $\Phi(R^2) = p$. The appearance of the χ^2_m should not be too surprising when we remember that this distribution applies to the sum of m squares of normally-distributed rv's. Using a standard table of the cdf Φ of χ^2, we find that, for $p = 0.95$, R is 1.96 for one dimension, 2.45 for two, and 2.79 for three. The $m = 1$ value is the result we had before: to have 0.95 probability, the limits are (about) 2σ from the mean. For higher dimensions, the limits become larger: as the pdf spreads out, a larger volume is needed to contain the same amount of probability. When we are (for example) plotting the two-dimensional equivalent of error bars we must take these larger limits into account.

4.7.1 The Multivariate Normal and Linear Regression

For a multivariate normal the regression functions are relatively simple; their form is widely used to model many regression problems. For a bivariate normal, we find one regression curve by taking the conditional pdf for X_2 with x_1 given:

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

This is a somewhat messy version of the univariate pdf; from it, we see that the expected value of X_2 is

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

which defines a **regression line** that passes through (μ_1, μ_2), and has a slope of $\rho(\sigma_1/\sigma_2)$ If we switch all the subscripts we get the regression line for X_1 given x_2:

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

which passes through the same point, but has a different slope: $\rho^{-1}(\sigma_1/\sigma_2)$. These regression lines are the same only for $\rho = 1$: perfect correlation; if ρ is zero, the regression lines are perpendicular. The lower panels of Figure 4.6 show the regression lines for two bivariate pdf's with correlation.

For $0 < \rho < 1$ neither line falls along the major axis of the elliptical contours, which causes **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. In the original example of Galton, because parental height and offspring height are imperfectly correlated 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 on average will have shorter parents. Repeated measurements of the same thing are often imperfectly correlated, in which case remeasuring items that were below the mean will, give values that are, on average, larger. When this happens, it is all too tempting to assume that this shift towards the mean implies a systematic change, rather than just random variations.

In higher dimensions the multivariate normal produces, not regression lines, but planes or hyperplanes; but the regression function of each variable on all the others is linear:

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

where, as in the bivariate case, the slopes a_j depend on the covariances, including of course the variances, So the multivariate normal pdf gives rise to **linear regression**, in which the relationships between variables are described by linear expressions. For the multivariate normal, having these linear expressions is almost enough to describe the pdf: they give its location and shape, with only one variance needed to set the spread of the pdf.

4.7.2 Linear Transformations of Multivariate Normal RV's

Suppose we form a linear combination of our original random variables $\vec{X}$ to produce a new set $\vec{Y}$; an example of such a transformation is summing random variables, each perhaps weighted by a different amount. In general, a linear transformation takes the original m rv's $\vec{X}$ and produces n rv's $\vec{Y}$; we can write this as

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

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 pdf of $\vec{X}$ is a multivariate normal, the pdf for $\vec{Y}$ can be shown to be another multivariate normal with mean value

$$\vec{\nu} = L\vec{\mu} \tag{4.1}$$

and covariance

$$C' = LCL^T \tag{4.2}$$

We demonstrate this in Section 4.8.

Two Examples: Weighted Means, and Sum and Difference One example of a linear combination of rv's 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 a Normal distribution and first and second moments μ and σ^2. It is immediately obvious from 4.1 that $\nu = \mu$; applying 4.2 to the covariance matrix C, which in this case is a scaled version of the identity matrix, we see that the variance of Y is

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

If all the weights are the same this gives the familiar result σ^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

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \tag{4.3}$$

The most general form for the bivariate covariance matrix C is

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{4.4}$$

If we now apply 4.2 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 if $\sigma_1 = \sigma_2 = \sigma$ 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 (equation 4.3) 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 one appropriate to two uncorrelated variables.

One application of this is 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 the sum and difference of the two, to remove this linear dependence so you can see other effects. The other application is in fitting two very similar functions to data. The results of the fit will appear to have large variances, which is correct; but there will also be a large correlation. Other combinations may have less correlation and more informative variances: the difference between functions may have a very large variance while the variance of the sum will be small.

4.7.3 Removing the Correlations from a Multivariate Normal

Taking sums and differences is an example of how, for a normal pdf, we rotate the coordinate axes to make the covariance matrix diagonal, and so create new and uncorrelated variables. For a multivariate normal pdf, this can be generalized: it is in fact always possible to produce a set of uncorrelated variables.

To derive this, we start by noting that any rotation of the axes involves transforming one set of variables to another using linear combinations, something we have already discussed. For a rotation the matrix of linear combinations, L, is square ($n = m$), and it and its inverse are orthogonal ($L^T = L^{-1} \overset{\text{def}}{=} K$ and $K^T = K^{-1}$). Equations (4.1) and (4.2) show that the mean and covariance of the new variables will be $\vec{v}$ and

$$C' = LCK = K^T CK \tag{4.5}$$

Zero correlations between the new variables means that C' must not have any nonzero covariances, and so must be diagonal:

$$K^T CK = \text{diag}[\sigma_1^2, \sigma_2^2, \dots \sigma_m^2]. \tag{4.6}$$

A symmetric positive definite matrix (which the covariance matrix C always is), can always be written as a product of a diagonal matrix and an orthogonal one:

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

where U is orthogonal, and Λ is diagonal; the values of the diagonal of Λ, $\lambda_1, \lambda_2, \dots \lambda_m$, are all positive. Then to get a diagonal covariance matrix we just take $K = U^T$, so that the transformation 4.5 produces the diagonal matrix Λ, so that the variances become $\sigma_j^2 = \lambda_j$. If we write the j-th column of U as $\vec{u}_j$, the orthogonality of U means that $\vec{u}_j \cdot \vec{u}_j = \delta_{jk}$: the columns of U are orthogonal. (These column vector are the eigenvectors of C, as the λ_j's are the eigenvalues). Thus, these column vectors $\vec{u}_j$ give a new set of directions, called the **principal axes**, for which the covariance matrix is diagonal and the variables are uncorrelated.

This rotation of axes to create uncorrelated variables has two uses, one unexceptionable and one dubious. It is unproblematic to provide the directions and variances as a compact summary of the full covariance matrix – indeed, if the covariances are nonzero, it is quite uninformative to provide only the variances. The dubious aspect comes from sorting the λ's from largest to smallest, and asserting that the directions associated with the largest σ's necessarily identify some meaningful quantity. Applying diagonalization and sorting to covariance matrices estimated from actual data, and so finding patterns of variation, is called **principal component analysis**; such an analysis usually shows that a few directions (that is, a few combinations of observations) account for most of the variation. Sometimes these few combinations are meaningful; but they may not be – that is another matter that calls for informed judgment, not mechanical application of a mathematical procedure.

4.8 Propagation of Errors

Often a multivariate normal adequately represents the pdf of some variables; even more often, we implicitly assume this by working only with the first two moments which completely specify a normal. So, suppose we consider only first moments (means) and second moments (variances and covariances) of a multivariate normal; and suppose further that we have a function $f(\vec{X})$ of the random variables; what multivariate normal should we use to approximate the pdf of this function?

This question is usually answered through the procedure known as the **propagation of errors**; in which we not only confine ourselves to the first and second moments, but also assume that we can adequately represent the function f by a linear approximation. This procedure is much simpler than what is necessary (Section 2.10) to find the exact pdf of one rv transformed from another, but this approximate and simpler procedure is very often adequate.

We start by assuming that the function $f(\vec{X})$ produces a single random variable. The linear approximation of the function is then

$$f(\vec{X}) = f(\vec{\mu}) + \sum_{i=1}^{m} \left(\frac{\partial f}{\partial x_i}\bigg|_{x_i=\mu_i} \right)(X_i - \mu_i) = f(\vec{\mu}) + \vec{d} \cdot (\vec{X} - \vec{\mu})^T \tag{4.7}$$

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 do not need to do integrals to find this expectation if we apply the linearity properties of the expectation operator $\mathscr{E}$ introduced in Section 2.9:

$$\mathscr{E}[f(\vec{X})] = f(\vec{\mu}) + \vec{d} \cdot \mathscr{E}[\vec{X} - \vec{\mu}]^T = f(\vec{\mu}) \tag{4.8}$$

because $\mathscr{E}(\vec{X}) = \vec{\mu}$. This is exact only if the linear relationship 4.7 is also exact.

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

$$\begin{aligned}\mathscr{V}(f(\vec{X})) =& \mathscr{E}\left[(\vec{d}\cdot(\vec{X}-\vec{\mu})^T)^2\right]\\ =&\mathscr{E}[(\vec{d}^T\cdot(\vec{X}-\vec{\mu}))\cdot(\vec{d}\cdot(\vec{X}-\vec{\mu})^T)]\\ =&\mathscr{E}[((\vec{X}-\vec{\mu})\cdot\vec{d}^T)\cdot(\vec{d}\cdot(\vec{X}-\vec{\mu})^T)]\\ =&\mathscr{E}[(\vec{X}-\vec{\mu})\cdot A\cdot(\vec{X}-\vec{\mu})^T]\end{aligned}$$

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

$$\begin{aligned}\mathscr{V}(f(\vec{X})) =& \mathscr{E}\left[\sum_i\sum_j a_{ij}(x_i-\mu_i)(x_j-\mu_j)\right]\\ =&\sum_i\sum_j a_{ij}\mathscr{E}[(x_i-\mu_i)(x_j-\mu_j)]\\ =&\sum_i\sum_j a_{ij}C_{ij} = \sum_{ij} d_i d_j C_{ij}\end{aligned}$$

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

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

So we have an expression for the variance of the function; this depends on both its partial 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:

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

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 $\vec{d}_b^T\cdot C\cdot\vec{d}_b$.

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

$$C' = DCD^T$$

which is of course of the same form as the transformation to a new set of axes given in Section 4.7.3 – though this result is approximate, while that was not. Along the diagonal, we get 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

A simple example is provided by the different sets of random variables that would describe a periodic phenomenon with period τ. This can be done in two ways

$$Y(t) = A\cos\left(\frac{2\pi t}{\tau} + P\right) \quad \text{or} \quad Y(t) = X_1 \cos\left(\frac{2\pi t}{\tau}\right) + X_2 \sin\left(\frac{2\pi t}{\tau}\right)$$

The random variables are either the amplitude A and phase P, or the cosine and sine amplitudes X_1 and X_2.[4] The first formula is more traditional, and easier to interpret; unlike the X's, the amplitude A does not depend on what time corresponds to $t = 0$. Using X_1 and X_2 is however preferable for discussing errors (and for other purposes), because the two parameters are linearly related to the result.

Using propagation of errors, how do the first and second moments of X_1 and X_2 map into the first and second moments of A and P? The relationship between these parameters is

$$A = f_1(X_1, X_2) = \sqrt{X_1^2 + X_2^2} \quad \text{and} \quad P = f_2(X_1, X_2) = \arctan\left(\frac{X_2}{X_1}\right) \qquad (4.9)$$

[4] X_1 and X_2 are also called the **in-phase** and **quadrature** parts; or, if a complex representation is used, the real and imaginary parts.

so the Jacobian matrix is

$$\begin{aligned} d_{11} = \tfrac{\partial f_1}{\partial x_1} &= x_1/a & d_{12} &= x_2/a \\ d_{21} = \tfrac{\partial f_2}{\partial x_1} &= -x_2/a^2 & d_{22} &= x_1/a^2 \end{aligned}$$

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 equation (4.4) we find that the variances for A and P are

$$\sigma_A^2 = \left(\frac{x_1}{a}\right)^2 \sigma_1^2 + \left(\frac{x_2}{a}\right)^2 \sigma_2^2 - \frac{x_1 x_2}{a^2}\rho\sigma_1\sigma_2 \qquad \sigma_P^2 = \frac{x_2^2}{a^4}\sigma_1^2 + \frac{x_1^2}{a^4}\sigma_2^2 - \frac{x_1 x_2}{a^4}\rho\sigma_1\sigma_2$$

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

$$\sigma_A = \sigma \qquad \sigma_P = \frac{\sigma}{a}$$

which makes sense if thought about geometrically: provided a is much larger than σ_1 and σ_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 and the result (4.8) no longer holds. First, the expected value of the amplitude is systematically larger than $(X_1^2 + X_2^2)^{1/2}$; in the limit, with the X's both zero, A is Rayleigh-distributed, and has a nonzero expected value. Second, the pdf for the angle P, is also not well approximated as a Normal. This is a particular example case of a more general result: if you have a choice between variables with normal errors and other variables that are 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

To transition to the next chapter on estimation, we return to the regression curve discussed in Section 4.6 and compare this with a superficially similar procedure that is often (confusingly) also called regression; while these are mathematically similar they are conceptually different.

In Section 4.6, we defined regression by starting with a pdf for at least two variables, and asking 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.

But "regression" is also used for fitting a partly-specified function to data. In this usage, 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. The regression curve (or more generally surface) is then the function we wish to find. But this situation lacks the symmetry we have in our earlier discussion, since not all the variables are random.

This procedure, which is much used in elementary discussions of data analysis, comes from a model we mentioned, and discarded, in Section 1.1.2. In this model (equation 1.3) we regard data y as being the sum of a function $f_p(\vec{x})$ and a random variable $\vec{E}$, where f_p depends on some parameters p and takes as argument some variables $\vec{x}$, and $\vec{E}$. In our terminology, $\vec{x}$ is a vector of conventional variables and f_p produces a conventional variable; only the error is modeled as a random variable. A simple example would be measuring the position of a falling object (with zero initial velocity) as a function of known times $\vec{t}$. We would write this as

$$\vec{Y} = \tfrac{1}{2}g\vec{t}^2 + \vec{E} \tag{4.10}$$

where $\vec{t}^2$ 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 $\phi(y)\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, so the individual Y's are independent and identically distributed. Then 4.10 can be written as a distribution for Y:

$$Y_i \sim \phi(y - \tfrac{1}{2}gt_i^2) \tag{4.11}$$

so that g, and the known values t_i, appear as parameters in a pdf – which, as conventional variables, is where they should 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 we study extensively in Chapter 5.

You should 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 one. However, we hope that it is also easy to see that we can engage in the activity of finding a curve that represents $\mathscr{E}(Y)$ 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 while we end up with similar solutions of these two problems, we assumed very different mathematical models to get them.