

Inverse Theory Weeks 1 and 2

H. C. Pumphrey

September 23, 2008

1 Introduction

In other courses (mostly in *Radiative Transfer* and *Fundamentals for remote sensing*) this semester we will learn how to calculate what a satellite instrument would detect if the ground and atmosphere at which it was looking were in a given state. This is called the *Forward Problem*. In this course, we will learn how to infer things about the surface and the atmosphere, given what an instrument has measured. This is called the *Inverse Problem*. Mathematically similar problems occur in other areas of Earth science (and science in general). In this course we use remote sensing of the atmosphere as an example but it should be borne in mind that the ideas and techniques we will examine have other uses.

In one sense, the Forward Problem is the hard part as it contains all the physics of the situation, while the Inverse Problem is just a mathematical exercise. In another sense, the Forward Problem is easy because, given all of the necessary physics, you put in the state of the atmosphere and there is one straightforward correct answer for the radiance arriving at the satellite. In many cases, the Inverse Problem is fundamentally under-constrained. By this we mean that the instrument does not provide enough information for us to specify exactly what the system being studied is like. The problem of sensing the temperature of the atmosphere is always like this, because the temperature is a continuous function of altitude and therefore contains infinitely many pieces of information, while the instrument can only measure a finite number.

Quite apart from the under-constrained-ness of the system, we have to contend with the fact that all real instruments suffer from some form of measurement noise – if you were able to make the same measurement several times over, you would not get exactly the same number each time. In these first two lectures, we look briefly at the nature of under-constrained problems to give you a feel for what is coming later. We then revise some of the the basic conceptual and mathematical tools you will need to understand under-constrained problems and experimental noise.

1.1 Recommended books

We recommend the following books for the whole of Remote Sensing 2.

1.1.1 Mathematics

- Mathematical Methods for the Physics and Engineering by Riley, Hobson and Bence.
- Advanced Engineering Mathematics by Erwin Kreysig.

Both these books provide all you need on matrices and at least some help with the statistics.

1.1.2 Atmosphere

- Remote Sounding of the Atmosphere, by J. Houghton, F. Taylor and C. D. Rodgers.
- Introduction to the Mathematics of Inversion in Remote Sensing, by S. Twomey.
- Inverse Methods for Atmospheric Sounding: Theory and Practise, by C. D. Rodgers.¹

2 Under-constrained problems.

Let's suppose we want to measure the sea surface temperature from space. If we forget that there is any atmosphere, we can calculate what radiance, at some infra-red wavelength or other, will arrive at our space ship. We will use a standard notation and call the thing we can measure y and the thing that we want to know x . Now we can write

$$y = f(x) \quad (1)$$

where the function f is the Forward Model and represents all the physics you will learn in Radiative Transfer. We'll assume for now that the problem is linear so that we can write :

$$y - y_0 = k(x - x_0)$$

where the constant k is a proportionality constant between what we want to know and what we can measure. We have linearised about x_0 which is a typical value for x ; y_0 is given by $y_0 = f(x_0)$. In this case, we can immediately write

$$(x - x_0) = k^{-1}(y - y_0)$$

¹This is easily the most relevant book for most of the course, although it covers the material in considerably more detail than the course does. The library have a copy and you can also borrow a copy from me. If you like it, you can order it from the bookshop. It is about 26.

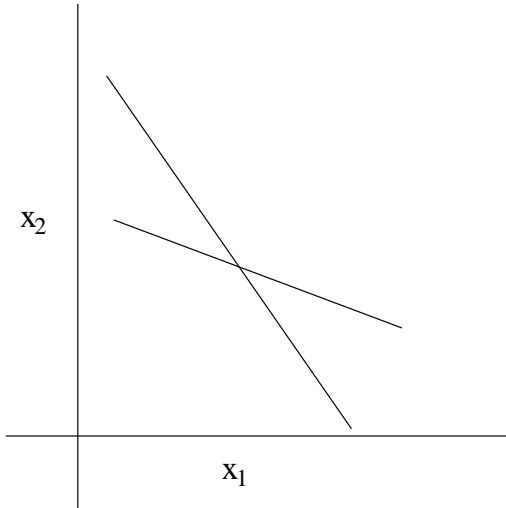


Figure 1: A well-posed problem with two equations in two unknowns. The solution is where the two lines cross.

to obtain the x that we require. From here on we shall assume that x_0 and y_0 are both zero in order to make things tidier.

Of course, we have made a huge simplification because in real life there is an atmosphere between the ocean and the spacecraft. The quantity we want to know therefore consists of two numbers: the sea surface temperature and the transmission through the atmosphere. We'll call these two numbers x_1 and x_2 . If we are to measure both of these we will need to measure two separate numbers from the space ship. (As an example, the ATSR instrument does this by looking at the same bit of the surface vertically, and at some different angle.) This provides us with two measurements which we will call y_1 and y_2 . The forward problem now looks like this:

$$y_1 = k_{11}x_1 + k_{12}x_2 \quad (2)$$

$$y_2 = k_{21}x_1 + k_{22}x_2 \quad (3)$$

Here, we have two simultaneous equations for two variables. They normally have one single correct answer; you probably learned several ways to find this answer at school. This is an example of a well-determined problem. Graphically, we can solve the problem by drawing a graph of x_1 against x_2 . Equations 2-3 become the two straight lines

$$x_2 = (y_1 - k_{11}x_1)/k_{12}$$

$$x_2 = (y_2 - k_{21}x_1)/k_{22}$$

which intersect at the values of x_1 and x_2 corresponding to the measured values y_1 and y_2 . This situation is sketched in Figure 1.

Suppose, however, that we had just one of the two measurements. We now have two unknowns and only one equation as suggested in Figure 2.

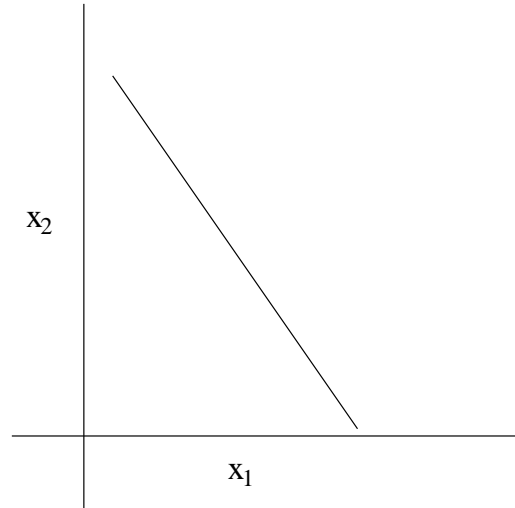


Figure 2: Two unknowns, but only one equation. The problem is under-constrained.

Now we do not have a unique solution. Our one measurement tells us that the solution cannot lie just anywhere in the (x_1, x_2) plane but is somewhere along the line. Without the second measurement, we don't know where along the line. This does not necessarily mean that the measurement is useless. We may be able to find some good reason apart from the measurements to say where along the line the solution might lie. This is called a *constraint*, figure 3 shows an example. In many cases getting useful results from remote sensing measurements depends on choosing sensible constraints to supplement the measurements.

In some cases, of course, we have more measurements than we have unknowns. This type of problem is known as an over-determined problem. In our example with two unknowns the problem would look like this

$$y_1 = k_{11}x_1 + k_{12}x_2 \quad (4)$$

$$y_2 = k_{11}x_1 + k_{12}x_2 \quad (5)$$

$$y_3 = k_{31}x_1 + k_{32}x_2 \quad (6)$$

and the graphical solution like figure 4. Note that in an ideal world, the third measurement tells us nothing that we didn't know already from the first two. In the real world it is more likely that there is no solution which agrees with all three measurements. The main reason for this is that there is usually some measurement error in the measurements y_i . Graphically, this makes the situation more like Figure 5; the lines now have some thickness and the true solution probably lies in the region where they cross. The addition of a third measurement may now be of some use as it may give a more accurate answer. The usual procedure in an over-constrained problem with noisy measurements is to call it a least-squares solution. We will cover this later in the course.

Since it is clear that experimental noise will be so important we will need to make use of the mathe-

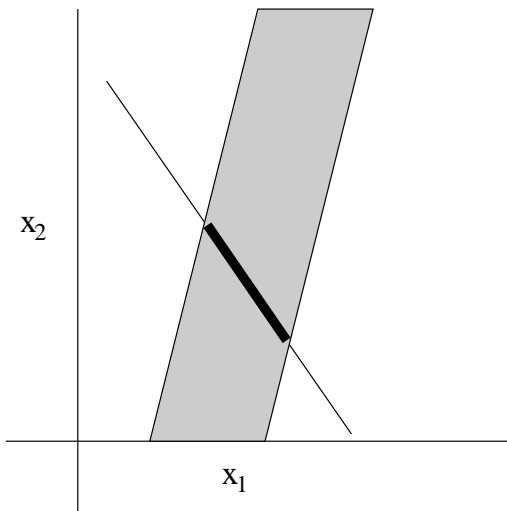


Figure 3: An under-constrained problem made more useful by an additional constraint. Our measurement tells us that the solution is along the line as before. However, we imagine that there is some physical reason why the solution must lie in the shaded area. Combined with the measurements, we now know that the solution is along the thick part of the line.

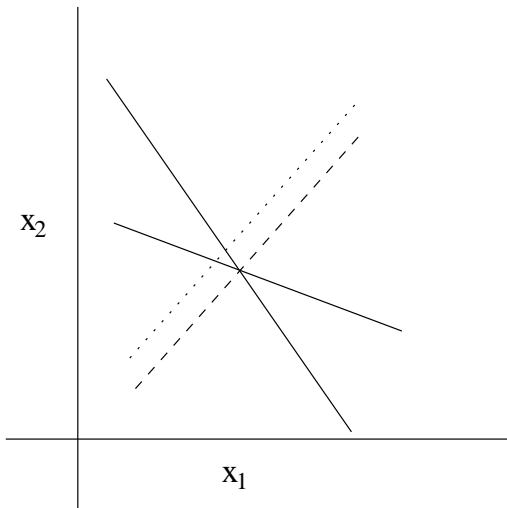


Figure 4: An over-determined problem. The third (dashed) line tells us nothing that we didn't know from the other two. If something is wrong with one of the lines (e.g. if our third measurement was indicated by the dotted line instead of the dashed one) then there is no unique solution.

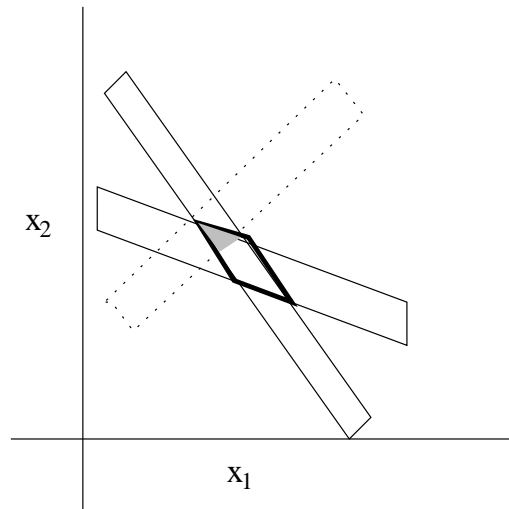


Figure 5: This figure shows a problem with two unknowns and where the measurements have some finite error. The first two measurements, shown by the plain lines, do not now meet at a single point but suggest a solution in the region marked by the thick line. The third (dotted) measurement narrows this down to the shaded area.

matics of random variables in order to study it further. This will be the topic of the second lecture. First, we provide a brief revision of all the things we need to know about matrices.

3 Matrices

3.1 Introduction

The aim of this section is to mention the aspects of matrix theory that you will need for the rest of the course. This will be revision material for most people. The parts that are not will perhaps go past too fast but will hopefully help you to read the right bits out of a textbook to bring you up to speed quickly.

3.2 Basics

A matrix is a rectangular array of numbers. The individual elements are referred to by two subscripts, like this:

$$\begin{matrix}
 A_{11} & A_{12} & A_{13} & \dots & A_{1n} \\
 A_{21} & A_{22} & A_{23} & \dots & A_{2n} \\
 A_{31} & A_{32} & A_{33} & \dots & A_{3n} \\
 \vdots & \vdots & \vdots & & \\
 A_{m1} & A_{m2} & A_{m3} & & A_{mn}
 \end{matrix}$$

We would call the example a $m \times n$ matrix. If $m = n$ it is called a square matrix. The special case where $m = 1$ or $n = 1$ is called a vector. I will attempt to stick to the following notation:

- Matrices are bold capitals like this: **A**

- Vectors are bold lower-case, like this: \mathbf{x}
- Scalars are italic, like this: s

We will also let the matrix \mathbf{A} have elements A_{ij} . You can add or subtract two matrices if they are the same size – $\mathbf{C} = \mathbf{A} + \mathbf{B}$ has elements $C_{ij} = A_{ij} + B_{ij}$. You can multiply an $m \times q$ matrix by a $q \times n$ matrix to get a $m \times n$ matrix: if $\mathbf{C} = \mathbf{AB}$ then $C_{ij} = \sum_{k=1}^q A_{ik}B_{kj}$. Two matrices which can be multiplied are said to be conformable. Matrix multiplication is not commutative: you cannot always calculate both \mathbf{AB} and \mathbf{BA} and when you can they are not necessarily equal.

If a matrix \mathbf{A} has elements A_{ij} , then its transpose, written \mathbf{A}^T , is a matrix with elements A_{ji} . We note that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. A matrix for which $\mathbf{A} = \mathbf{A}^T$ (i.e. for which $A_{ij} = A_{ji}$) is said to be symmetric. Here are some other types of special matrices we will encounter.

- A diagonal matrix is a square matrix for which $A_{ij} = 0 \forall i \neq j$.
- A unit matrix (written \mathbf{I}) is a diagonal matrix with $A_{ii} = 1 \forall i$. $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$ for any matrix \mathbf{A} which is conformable with \mathbf{I} .

3.3 The inverse matrix and simultaneous equations

For a square matrix \mathbf{A} there may or may not exist a matrix \mathbf{A}^{-1} such that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. This matrix is called the inverse of \mathbf{A} . If \mathbf{A}^{-1} does not exist, then \mathbf{A} is said to be singular. The set of simultaneous equations

$$\begin{aligned} y_1 &= k_{11}x_1 + k_{12}x_2 \\ y_2 &= k_{21}x_1 + k_{22}x_2 \end{aligned}$$

may be written more concisely as $\mathbf{y} = \mathbf{Kx}$. The solution, if it exists, can therefore be written as $\mathbf{x} = \mathbf{K}^{-1}\mathbf{y}$. If the solution does not exist then \mathbf{K} is a singular matrix. Note that finding the inverse of a matrix is not the most efficient way to solve for \mathbf{x} but that writing the equations in this form is often conceptually useful.

3.4 Determinants

A determinant is a single number that can be calculated for a square matrix. The determinant of \mathbf{A} is written $|\mathbf{A}|$. For a 2×2 matrix, $|\mathbf{A}|$ is defined as

$$|\mathbf{A}| = \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = A_{11}A_{22} - A_{12}A_{21}$$

For a 3×3 matrix, $|\mathbf{A}|$ is defined as

$$|\mathbf{A}| = A_{11} \begin{vmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{vmatrix} - A_{12} \begin{vmatrix} A_{21} & A_{23} \\ A_{31} & A_{33} \end{vmatrix} + A_{13} \begin{vmatrix} A_{21} & A_{22} \\ A_{31} & A_{32} \end{vmatrix}$$

The three 2×2 determinants are called the minors of the three elements A_{11} , A_{12} and A_{13} – you get the minor of an element by deleting the row and column of the original determinant that contain that element. Determinants for larger matrices are defined in a similar way.

One way of calculating the inverse of a matrix is from the formula $\mathbf{A}^{-1} = \text{adj}(\mathbf{A})/|\mathbf{A}|$. We won't worry about how to find the adjoint matrix $\text{adj}(\mathbf{A})$ as this is not a computationally efficient way to find an inverse. We present the formula because it allows us to see that if the determinant of a matrix is zero, then the inverse will not exist, i.e. \mathbf{A} will be singular.

3.5 Simultaneous equations again.

Consider again the set of simultaneous equations $\mathbf{y} = \mathbf{Kx}$, where \mathbf{K} is a $m \times n$ matrix. We saw already that if \mathbf{K} is not square i.e. if $m \neq n$, then there is generally no unique solution.

If $m < n$ then the problem is under-determined and there are infinitely many solutions. It is sometimes useful to obtain one of these solutions i.e. any vector \mathbf{x} for which $\mathbf{y} = \mathbf{Kx}$. To do this, we need an $n \times m$ matrix \mathbf{D} for which $\mathbf{DK} = \mathbf{I}$. Multiply on the right by \mathbf{K}^T to give $\mathbf{DKK}^T = \mathbf{K}^T$. Now, \mathbf{KK}^T is a $m \times m$ matrix which we can probably take the inverse of. We can therefore write $\mathbf{D} = \mathbf{D}(\mathbf{KK}^T)(\mathbf{KK}^T)^{-1} = \mathbf{K}^T(\mathbf{KK}^T)^{-1}$.

If $m > n$ we have the opposite situation. We have more equations than unknowns and there may be no solution that satisfies all of them. There is no \mathbf{x} such that $\mathbf{y} = \mathbf{Kx}$ exactly. In this situation, it is often helpful to find the \mathbf{x} which makes \mathbf{Kx} as close to \mathbf{y} as possible. Since $\mathbf{y} - \mathbf{Kx}$ is a vector, we make its squared length $C = (\mathbf{y} - \mathbf{Kx})^T(\mathbf{y} - \mathbf{Kx})$ as small as possible – this is called a least squares solution. To find the \mathbf{x} that makes C a minimum, we do exactly what we would do if \mathbf{x} was a scalar – we differentiate with respect to \mathbf{x} and set the result to zero. The only difference is that $\frac{dC}{d\mathbf{x}}$ is a vector with elements $\frac{dC}{dx_i}$. Expanding out C gives:

$$C = \mathbf{y}^T\mathbf{y} + \mathbf{x}^T\mathbf{K}^T\mathbf{Kx} - \mathbf{y}^T\mathbf{Kx} - \mathbf{x}^T\mathbf{K}^T\mathbf{y}$$

Differentiating a scalar expression like this with respect to a vector is fairly simple, but the result often ends up a mix of row and column vectors: you have to re-arrange some of the terms so they are all column vectors. Here, the result is:

$$\frac{dC}{d\mathbf{x}} = 2\mathbf{K}^T\mathbf{Kx} - \mathbf{y}^T\mathbf{K} - \mathbf{K}^T\mathbf{y}$$

but the middle term is a row vector. Transposing it to get a column vector like the other terms gives:

$$\frac{dC}{d\mathbf{x}} = 2\mathbf{K}^T\mathbf{Kx} - 2\mathbf{K}^T\mathbf{y}$$

Setting this equal to zero gives $\mathbf{K}^T\mathbf{Kx} = \mathbf{K}^T\mathbf{y}$ and hence:

$$\mathbf{x} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{y}$$

3.6 Eigenvalues and eigenvectors

A square matrix \mathbf{A} has a set of properties called its eigenvalues and eigenvectors. An eigenvector of \mathbf{A} is a vector \mathbf{x} such that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, i.e. it is a vector which, when multiplied by \mathbf{A} gives a result which points in the same direction. The number λ is the eigenvalue associated with the eigenvector. You can find the eigenvalues and eigenvectors of a matrix like this: if $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ then $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$. That is a set of simultaneous equations: if $(\mathbf{A} - \lambda\mathbf{I})^{-1}$ exists, then we can multiply by it to get the dull answer $\mathbf{x} = 0$. For any more interesting answers to exist, we need $(\mathbf{A} - \lambda\mathbf{I})$ to have no inverse i.e. for $|\mathbf{A} - \lambda\mathbf{I}| = 0$. The determinant can be expanded to give a polynomial equation for λ - if \mathbf{A} was a $n \times n$ matrix then the equation will be of order n , giving n solutions λ_i for λ . These can be back-substituted in turn into the original equation to find n different eigenvectors \mathbf{x}_i . The eigenvectors are of an arbitrary magnitude: if you multiply one by any scalar it is still an eigenvector. It is common (and for some applications necessary) to normalise the eigenvectors, i.e. to make them so that $\mathbf{x}^T\mathbf{x} = 1$.

If \mathbf{A} is symmetric (as many of the matrices we meet later in the course are) then its eigenvalues and eigenvectors have two important properties:

- The eigenvalues are all real.
- The eigenvectors are orthogonal i.e. $\mathbf{x}_i^T\mathbf{x}_j = 0 \forall i \neq j$.

If we form a square matrix \mathbf{U} whose columns are the normalised eigenvectors, then $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. In other words, \mathbf{U} is a matrix whose transpose is its inverse: such a matrix is called an orthogonal matrix. We can write the original eigenvalue equation as $\mathbf{A}\mathbf{U} = \mathbf{U}\Lambda$, where Λ is a diagonal matrix with $\Lambda_{ii} = \lambda_i$. Hence, we can write the original symmetric matrix \mathbf{A} in terms of its eigenvectors: $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T = \sum_i \lambda_i \mathbf{x}_i \mathbf{x}_i^T$. This is called a spectral decomposition of \mathbf{A} .

4 Random variables

4.1 Scalar random variables

Suppose that an atmospheric physicist weighs himself three times in the morning, before breakfast. He will probably get three slightly different answers on account of the behaviour of the balance and how he goes about reading the scale. The measurement of his weight is an example of a *random variable*. Let's consider a random variable which we will call x . Individual attempts to find a value for x are called *samples* - we will indicate these by subscripts. If you weigh yourself three times you have three samples of the random variable that is your weight. To get a better estimate of x you would probably con-

sider finding the average or *mean*, \bar{x} , given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (7)$$

Another important thing to know is how scattered the measurements are about the mean. A commonly used measure of this is the Variance $V(x)$, given by:

$$V(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (8)$$

This gives the variance of your set of measurements. If the measurements are only part of a larger population of possible measurements, then the variance of the whole population is given by:

$$V_p(x) = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (9)$$

This is because the sample is not likely to sample enough of the extreme outliers of the full population. The standard deviation σ is just the square root of the variance. It is intuitively more useful than the variance itself as it has the same units or dimensions as x . If you use the "standard deviation" button on a calculator, do make sure that you know which of the two formulae 8 and 9 your calculator is using.

Often, we are concerned with measurements of two different variables x and z which are measured in pairs so that each measurement of x , x_i has a corresponding measurement of z , z_i . For example, x might be the weight of our atmospheric physicist and z his waistline. We can calculate variances for x and z separately, but we can also calculate a quantity called the *co-variance* of x and z , $V(x, z)$, given by

$$V(x, z) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z}) \quad (10)$$

The covariance measures how much the variations in x depend on those in z . If our fearless experimenter makes ten pairs of measurements of his weight and his waistline within a few minutes then the variations from the mean are all due to measurement error. There is no connection between the tape measure and the balance so the co-variance will be small. If, however, he makes one pair of measurements every day over the Christmas holidays, much of the variation might be due to his real weight and waistline changing. In this case, the covariance of the two quantities will be positive because as his weight goes up, so does his waistline. Note that while a variance is always positive, a co-variance can be negative. This will happen if x tends to be smaller than usual in samples where z is larger than usual.

4.2 Vector Random Variables

Some of the things we measure are scalars, some are vectors. The air temperature on the roof of this building is a scalar, the wind velocity is a vector. Suppose we have an m -dimensional vector quantity \mathbf{v} , with components

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \dots \\ v_m \end{bmatrix}$$

A single measurement of \mathbf{v} will be written as

$$\mathbf{v}_i = \begin{bmatrix} v_{1i} \\ v_{2i} \\ \dots \\ v_{mi} \end{bmatrix}$$

We can calculate a mean value for \mathbf{v} just as we did for a scalar variable. The “variance” of a vector random variable is a bit more complicated. We define \mathbf{S} , the co-variance matrix of \mathbf{v} as

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{v}_i - \bar{\mathbf{v}})(\mathbf{v}_i - \bar{\mathbf{v}})^T \quad (11)$$

Notice that the diagonal elements \mathbf{S}_{ii} are the variances of the individual elements of \mathbf{v} while an off-diagonal element \mathbf{S}_{ij} is the covariance of v_i and v_j .

4.3 Errors of combined measurements

Equations 7-9 tell us how to calculate the mean and variance of a set of measurement. We now ask: If we make a n measurements, how good an estimate of the true mean of our random variable is the result of equation 7? It is a standard result that the error in the mean is given by σ/\sqrt{n} , so that if you take the mean of 100 samples, the error in the mean will be 10 times smaller than the error in one sample. Taking this question a little further, we now ask how we combine two estimates of a quantity which are known to have different errors. Suppose we have two estimates of a quantity x : x_1 and x_2 . We suppose that these estimates are wrong by amounts ϵ_1 and ϵ_2 . We don’t know these exactly or we would know x exactly, but we do know that they have zero mean and standard deviations σ_1 and σ_2 . Now we try to form a \hat{x} , linear combination of x_1 and x_2 that will be a better estimate of x . We let

$$\hat{x} = ax_1 + bx_2$$

and calculate the error ϵ in \hat{x} :

$$\begin{aligned} \epsilon &= \hat{x} - x = ax_1 + bx_2 - x \\ &= a(x + \epsilon_1) + b(x + \epsilon_2) - x \\ &= x(a + b - 1) + a\epsilon_1 + b\epsilon_2 \end{aligned}$$

The error in \hat{x} should not depend on x – this would be a systematic error. We therefore require $a + b = 1$. We therefore have:

$$\epsilon = a\epsilon_1 + (1 - a)\epsilon_2 \quad (12)$$

We don’t know any of the epsilons, only their variances. It is a standard result in statistics that for two uncorellated random variables the variance of a sum is the sum of the variances². Applying this to equation 12 gives:

$$\sigma^2 = a^2\sigma_1^2 + (1 - a)^2\sigma_2^2$$

We want the value of a that makes this a minimum, so we differentiate with respect to a and set the result to zero, giving $a = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$ and hence:

$$\begin{aligned} \hat{x} &= \frac{(\sigma_2^2 x_1 + \sigma_1^2 x_2)}{(\sigma_1^2 + \sigma_2^2)} \\ \sigma^2 &= \frac{\sigma_1^2 \sigma_2^2}{(\sigma_1^2 + \sigma_2^2)} \end{aligned}$$

If we write S for the variance, so $S = \sigma^2$, then we can re-write these equations like this:

$$\begin{aligned} \hat{x} &= (S_1^{-1} + S_2^{-1})^{-1} (S_1^{-1} x_1 + S_2^{-1} x_2) \\ S &= (S_1^{-1} + S_2^{-1})^{-1} \end{aligned}$$

This makes it clearer that \hat{x} is a weighted mean of the two estimates and that the weights are the inverses of the individual variances. These equations work for a vector random variable, too, if you replace the variances by covariance matrices. The $^{-1}$ would then mean a matrix inverse.

4.4 Probability Density Functions

In this section we consider how likely it is that a random variable is going to take on a particular value when we measure it. We’ll make extensive use of the concept of probability. This is a number between 0 and 1 describing how likely it is that something will happen: 0 means it certainly won’t happen and 1 means it certainly will. Before you roll a die, for example, the probability that you will get a four is $\frac{1}{6}$. We write this as $p(4) = \frac{1}{6}$. The result of rolling a die is just another example of a random variable but it is different from the ones we have considered so far in that it can only take on six possible values: we call this a discrete random variable. A variable such as temperature which can take on any value (in a physically reasonable range) is called a continuous random variable. We can represent the probabilities of the various values of a discrete random variable as a bar chart; Figure 6 shows two examples.

We need to extend the concept of probability to continuous random variables. To do this we define the Probability Density Function $P(x)$ for the random variable x . This is a function of x such that the

²You can prove this by applying equation 8 to the sum of two random variables.

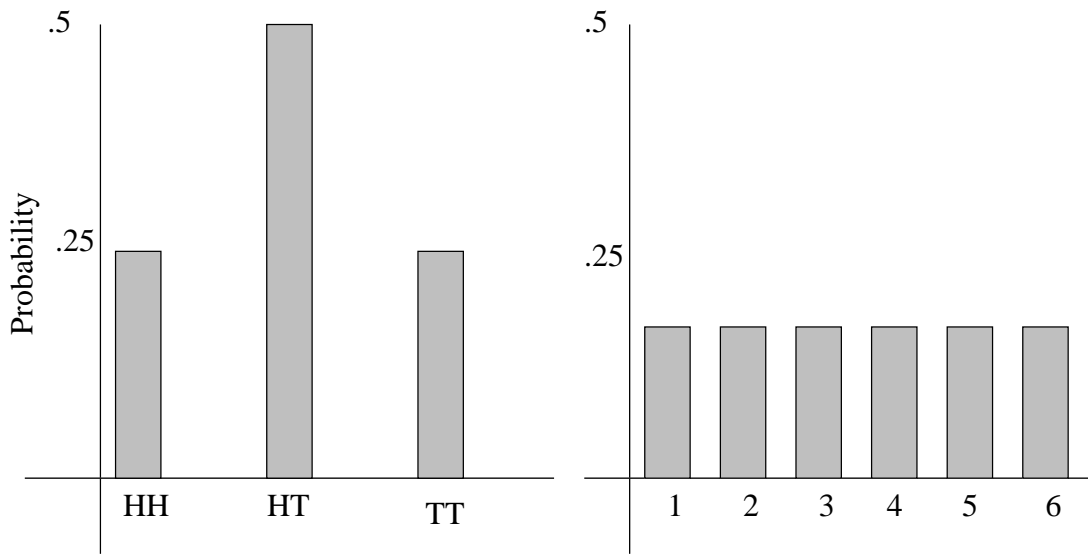


Figure 6: The right-hand plot shows the probabilities of the various outcomes of rolling a die. The left-hand plot shows the probabilities of the three possible outcomes of tossing two identical coins.

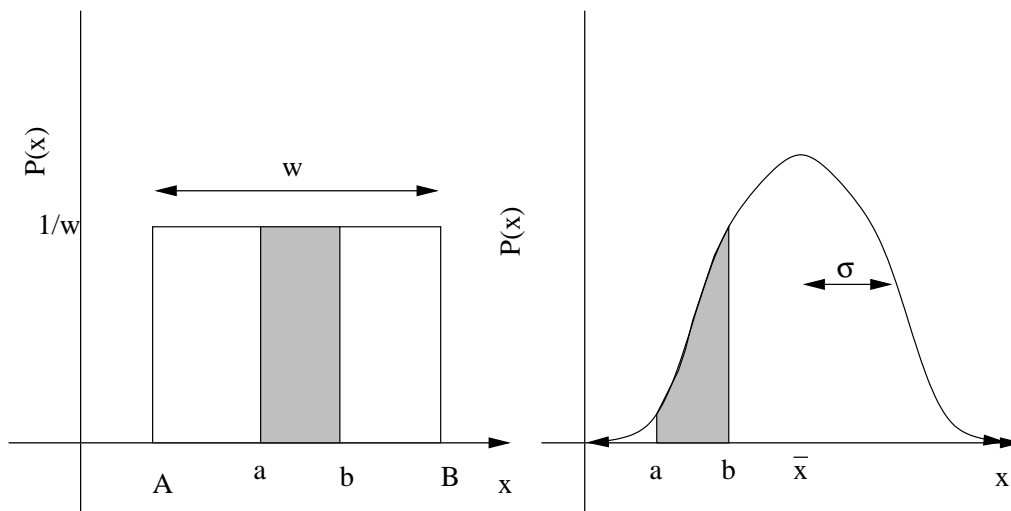


Figure 7: Two examples of what a probability density function might look like. In both cases, the probability that x_i will lie between a and b is given by the shaded area. The total area under each curve must be 1 as this is the probability that x_i will lie between $-\infty$ and $+\infty$. In the left-hand example, x_i is equally likely to take on any value between A and B and never lies outside this range. The right hand example is a Gaussian with a mean of \bar{x} and a standard deviation of σ .

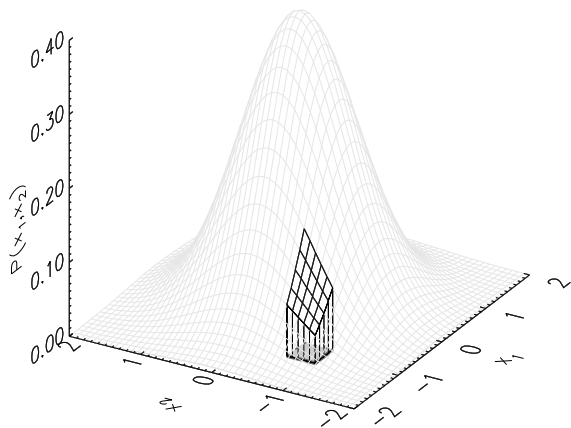


Figure 8: Gaussian PDF for a random vector $\mathbf{x} = (x_1, x_2)$ which has a mean of $(0, 0)$. The covariance matrix \mathbf{S} is a unit matrix. The probability that \mathbf{x} lies in the grey area is given by the volume outlined in black.

probability that a sample x_i will lie between x and $x + dx$ is $P(x)dx$. The probability that x_i will lie between two values a and b is given by $\int_a^b P(x)dx$. Figure 7 shows two examples of what a probability density function might look like. For reasons which are too deep to go into here, many random variables have a probability density function which can be reasonably well approximated by a Gaussian, or Normal distribution. This has the form:

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \bar{x})^2}{2\sigma^2}\right)$$

The concept of probability density functions extends easily from scalar random variables to vector ones. For brevity we will consider a vector random variable \mathbf{x} with two components x_1 and x_2 . We can then define $P(\mathbf{x})$ so that the probability that the two components of \mathbf{x} lie between x_1 and $x_1 + dx_1$ and between x_2 and $x_2 + dx_2$ is $P(\mathbf{x})dx_1dx_2$. The Gaussian distribution also has a vector form:

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{S}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}})\right)$$

Figure 8 shows an example of a Gaussian probability density function for a random vector with two elements. It is worth examining the appearance of a two-dimensional Gaussian PDF for various values of the covariance matrix \mathbf{S} . These are shown in Figures 9 to 12.

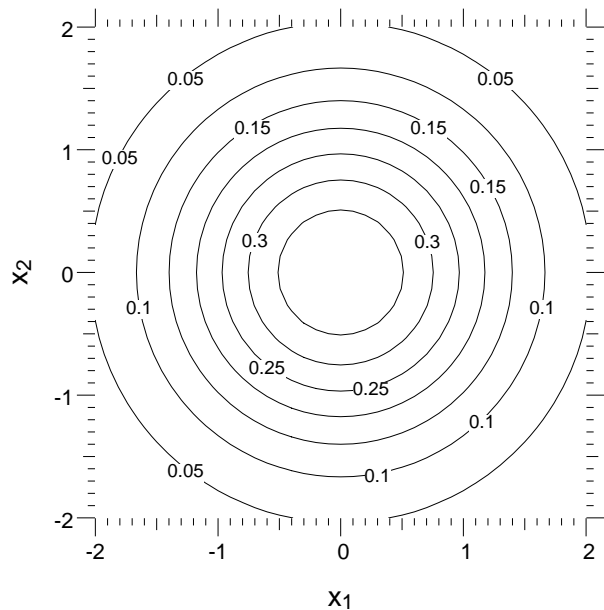


Figure 9: Gaussian PDF for a unit covariance matrix $\mathbf{S} = \mathbf{I}$.

Figure	\mathbf{S}	
9	1	0
	0	1
10	1	0
	0	0.1
11	1	0.8
	0.8	1
12	1	-0.8
	-0.8	1

4.5 Bayes' Theorem

We will now consider an important result which helps us to connect what we know about something before we make a measurement with what we know about it afterwards. We shall use the notation that $p(a)$ is the probability that the proposition a is true. For example, if a is the proposition that we will get a five if we roll a die, then $p(a) = \frac{1}{6}$. We also define two other, more complicated types of probability. Firstly, we define the *joint probability* of two propositions a and b , which is written $p(a, b)$ and is the probability that both a and b are true. Secondly, we consider the probability that a is true on the assumption that we know already that b is true. This is called the *conditional probability of a given b* and is written $p(a|b)$.

An important relation states that:

$$p(a, b) = p(a|b)p(b)$$

This may become clearer if you state it in words: The probability that a and b are both true is equal to the probability that a is true given that b is also true multiplied by the probability that b is true. From this we have:

$$p(b|a)p(a) = p(b, a) = p(a, b) = p(a|b)p(b)$$

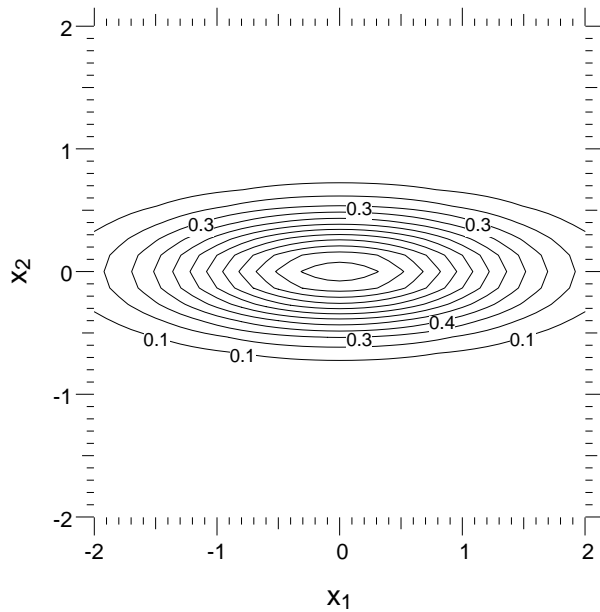


Figure 10: Gaussian PDF for diagonal (non-unit) covariance matrix.

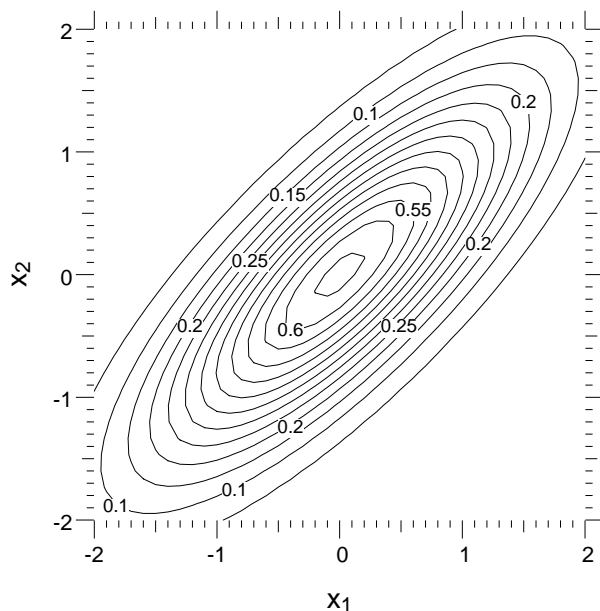


Figure 11: Gaussian PDF: Shas +ve off-diagonal elements

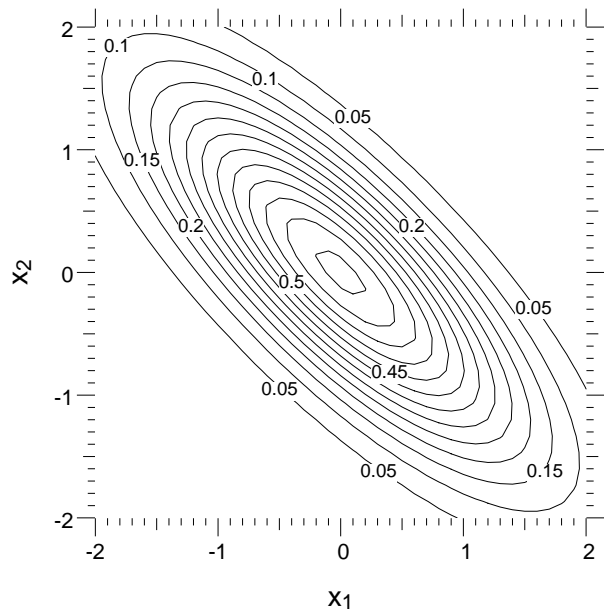


Figure 12: Gaussian PDF: Shas -ve off-diagonal elements

and hence

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)} \quad (13)$$

This is called Bayes' Theorem - it will be used extensively later in the course. This may look like rather an academic and pointless exercise but it has a rather general applicability in remote sensing for the following reason. Consider a to be what we can measure and b to be what we want to know, in some general sense. We can calculate $p(a|b)$ from the physics of the situation - the forward problem you are learning about in *radiative transfer*. Bayes' theorem then allows us to infer the PDF of the thing we want from the probability of the thing we can measure.

Here is a further example of the Bayes' theorem and the surrounding notation in action. Suppose a manufacturer of toy balls makes them in three sizes: small, medium and large, and in four colours: red, yellow green and blue. A particular shipment contains 1800 balls with the sizes and colours shown in the table.

	red	yellow	green	blue	total
small	50	100	100	150	400
medium	50	150	200	200	600
large	100	300	250	150	800
total	200	550	550	500	1800

We will use s for the proposition that a ball taken at random is small, g for the proposition that it is green, etc. So we have $p(l) = p(\text{it is a large ball}) = \frac{\text{number of large balls}}{\text{total number of balls}} = \frac{800}{1800} = \frac{4}{9}$. Similarly, the probability that a ball taken at random is yellow, $p(y)$, is given by $p(y) = \frac{550}{1800} = \frac{11}{36}$. The

probability that a ball is both yellow *and* large, $p(y,l)$, is $p(y,l) = 300/1800 = 1/6$. Suppose now we pick a ball at random chosen from the yellow balls only. The probability that it is a large ball given that it is yellow, $p(l|y)$ is given by $p(l|y) = 300/550 = 6/11$, the number of large, yellow balls divided by the total number of yellow balls. We can see that the relationship $p(y,l) = p(l|y)p(y) = 1/6 = (6/11) \times (11/36)$ is true for these numbers.

Now, suppose we draw a ball from the entire stock. The probability that it is yellow is $p(y)$. If we now feel the ball and determine that it is large, the probability that it is yellow changes. It is now no longer $p(y)$ but $p(y|l)$. In this case, we can deduce $p(y|l)$ directly from the table. Bayes' theorem enables us to deduce $p(y|l)$ in a situation where we knew $p(l|y)$: $p(y|l) = p(l|y)p(y)/p(l) = (6/11) \times (11/36)/(4/9) = 3/8$. You can check from the table that this is true.

Bayes' theorem can be applied to continuous random variables as well as to discrete ones. Suppose we have two random variables, x and y , with probability density functions $P(x)$ and $P(y)$. We define a joint probability density function $P(x, y)$, such that $P(x, y)dx dy$ is the probability that a sample of x lies between x and $x + dx$ *and* that a sample of y lies between y and $y + dy$. We also define a conditional probability density function $P(x|y)$, which is the probability density function for x given a particular value of y . As with the discrete case it is true that

$$P(x, y) = P(x|y)P(y)$$

and that

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