

## CHAPTER 2: REVIEW OF LINEAR ALGEBRA AND STATISTICS

### 2.1 Introduction

In discrete inverse methods, matrices and linear transformations play fundamental roles. So do probability and statistics. This review chapter, then, is divided into two parts. In the first, we will begin by reviewing the basics of matrix manipulations. Then we will introduce some special types of matrices (Hermitian, orthogonal and semiorthogonal). Finally, we will look at matrices as linear transformations that can operate on vectors of one dimension and return a vector of another dimension. In the second section, we will review some elementary probability and statistics, with emphasis on Gaussian statistics. The material in the first section will be particularly useful in later chapters when we cover eigenvalue problems, and methods based on the length of vectors. The material in the second section will be very useful when we consider the nature of noise in the data and when we consider the maximum likelihood inverse.

### 2.2 Matrices and Linear Transformations

Recall from the first chapter that, by convention, vectors will be denoted by lower case letters in boldface (i.e., the data vector  $\mathbf{d}$ ), while matrices will be denoted by upper case letters in boldface (i.e., the matrix  $\mathbf{G}$ ) in these notes.

#### 2.2.1 Review of Matrix Manipulations

##### *Matrix Multiplication*

If  $\mathbf{A}$  is an  $N \times M$  matrix (as in  $N$  rows by  $M$  columns), and  $\mathbf{B}$  is an  $M \times L$  matrix, we write the  $N \times L$  product  $\mathbf{C}$  of  $\mathbf{A}$  and  $\mathbf{B}$ , as

$$\mathbf{C} = \mathbf{AB} \quad (2.1)$$

We note that matrix multiplication is associative, that is

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) \quad (2.2)$$

but in general is not commutative. That is, in general

$$\mathbf{AB} \neq \mathbf{BA} \quad (2.3)$$

In fact, if  $\mathbf{AB}$  exists, then the product  $\mathbf{BA}$  only exists if  $\mathbf{A}$  and  $\mathbf{B}$  are square.

In Equation (2.1) above, the  $ij$ th entry in  $\mathbf{C}$  is the product of the  $i$ th row of  $\mathbf{A}$  and the  $j$ th column of  $\mathbf{B}$ . Computationally, it is given by

$$c_{ij} = \sum_{k=1}^M a_{ik} b_{kj} \quad (2.4)$$

One way to form  $\mathbf{C}$  using standard FORTRAN code would be

```

DO 300 I = 1, N
DO 300 J = 1, L
C(I,J) = 0.0
DO 300 K = 1, M
300 C(I,J) = C(I,J) + A(I,K)*B(K,J)
    
```

(2.5)

A special case of the general rule above is the multiplication of a matrix  $\mathbf{G}$  ( $N \times M$ ) and a vector  $\mathbf{m}$  ( $M \times 1$ ):

$$\mathbf{d} = \mathbf{G} \mathbf{m} \quad (1.13)$$

$(N \times 1) \quad (N \times M) \quad (M \times 1)$

In terms of computation, the vector  $\mathbf{d}$  is given by

$$d_i = \sum_{j=1}^M G_{ij} m_j \quad (2.6)$$

### *The Inverse of a Matrix*

The mathematical inverse of the  $M \times M$  matrix  $\mathbf{A}$ , denoted  $\mathbf{A}^{-1}$ , is defined such that:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_M \quad (2.7)$$

where  $\mathbf{I}_M$  is the  $M \times M$  identity matrix given by:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \quad (2.8)$$

$(M \times M)$

$\mathbf{A}^{-1}$  is the matrix, which when either pre- or postmultiplied by  $\mathbf{A}$ , returns the identity matrix. Clearly, since only square matrices can both pre- and postmultiply each other, the mathematical inverse of a matrix only exists for square matrices.

A useful theorem follows concerning the inverse of a product of matrices:

**Theorem:** If 
$$\mathbf{A} = \mathbf{B} \mathbf{C} \mathbf{D} \tag{2.9}$$

$$N \times N \quad N \times N \quad N \times N \quad N \times N$$

Then  $\mathbf{A}^{-1}$ , if it exists, is given by

$$\mathbf{A}^{-1} = \mathbf{D}^{-1} \mathbf{C}^{-1} \mathbf{B}^{-1} \tag{2.10}$$

**Proof:** 
$$\begin{aligned} \mathbf{A}(\mathbf{A}^{-1}) &= \mathbf{BCD}(\mathbf{D}^{-1}\mathbf{C}^{-1}\mathbf{B}^{-1}) \\ &= \mathbf{BC} (\mathbf{DD}^{-1}) \mathbf{C}^{-1}\mathbf{B}^{-1} \\ &= \mathbf{BC} \mathbf{I} \mathbf{C}^{-1}\mathbf{B}^{-1} \\ &= \mathbf{B} (\mathbf{CC}^{-1}) \mathbf{B}^{-1} \\ &= \mathbf{BB}^{-1} \\ &= \mathbf{I} \end{aligned} \tag{2.11}$$

Similarly,  $(\mathbf{A}^{-1})\mathbf{A} = \mathbf{D}^{-1}\mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{BCD} = \dots = \mathbf{I}$  (Q.E.D.)

*The Transpose and Trace of a Matrix*

The transpose of a matrix  $\mathbf{A}$  is written as  $\mathbf{A}^T$  and is given by

$$(\mathbf{A}^T)_{ij} = \mathbf{A}_{ji} \tag{2.12}$$

That is, you interchange rows and columns.

The transpose of a product of matrices is the product of the transposes, in reverse order. That is,

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \tag{2.13}$$

Just about everything we do with real matrices  $\mathbf{A}$  has an analog for complex matrices. In the complex case, wherever the transpose of a matrix occurs, it is replaced by the complex conjugate transpose of the matrix, denoted  $\tilde{\mathbf{A}}$ . That is,

$$\text{if } \mathbf{A}_{ij} = a_{ij} + b_{ij}\mathbf{i} \quad (2.14)$$

$$\text{then } \tilde{\mathbf{A}}_{ij} = c_{ij} + d_{ij}\mathbf{i} \quad (2.15)$$

$$\text{where } c_{ij} = a_{ji} \quad (2.16)$$

$$\text{and } d_{ij} = -b_{ji} \quad (2.17)$$

$$\text{that is, } \tilde{\mathbf{A}}_{ij} = a_{ji} - b_{ji}\mathbf{i} \quad (2.18)$$

Finally, the trace of  $\mathbf{A}$  is given by

$$\text{trace}(\mathbf{A}) = \sum_{i=1}^M a_{ii} \quad (2.19)$$

### *Hermitian Matrices*

A matrix  $\mathbf{A}$  is said to be Hermitian if it is equal to its complex conjugate transpose. That is, if

$$\mathbf{A} = \tilde{\mathbf{A}} \quad (2.20)$$

If  $\mathbf{A}$  is a real matrix, this is equivalent to

$$\mathbf{A} = \mathbf{A}^T \quad (2.21)$$

This implies that  $\mathbf{A}$  must be square. The reason that Hermitian matrices will be important is that they have only real eigenvalues. We will take advantage of this many times when we consider eigenvalue and shifted eigenvalue problems later.

## **2.2.2 Matrix Transformations**

### *Linear Transformations*

A matrix equation can be thought of as a linear transformation. Consider, for example, the original matrix equation:

$$\mathbf{d} = \mathbf{Gm} \quad (1.13)$$

where  $\mathbf{d}$  is an  $N \times 1$  vector,  $\mathbf{m}$  is an  $M \times 1$  vector, and  $\mathbf{G}$  is an  $N \times M$  matrix. The matrix  $\mathbf{G}$  can be thought of as an operator that operates on an  $M$ -dimensional vector  $\mathbf{m}$  and returns an  $N$ -dimensional vector  $\mathbf{d}$ .

Equation (1.13) represents an explicit, linear relationship between the data and model parameters. The operator  $\mathbf{G}$ , in this case, is said to be linear because if  $\mathbf{m}$  is doubled, for example, so is  $\mathbf{d}$ . Mathematically, one says that  $\mathbf{G}$  is a linear operator if the following is true:

$$\begin{aligned} \text{If } \quad \mathbf{d} &= \mathbf{G}\mathbf{m} \\ \text{and } \quad \mathbf{f} &= \mathbf{G}\mathbf{r} \\ \text{then } \quad [\mathbf{d} + \mathbf{f}] &= \mathbf{G}[\mathbf{m} + \mathbf{r}] \end{aligned} \tag{2.22}$$

In another way to look at matrix multiplications, in the by-now-familiar Equation (1.13),

$$\mathbf{d} = \mathbf{G}\mathbf{m} \tag{1.13}$$

the column vector  $\mathbf{d}$  can be thought of as a weighted sum of the *columns* of  $\mathbf{G}$ , with the weighting factors being the elements in  $\mathbf{m}$ . That is,

$$\mathbf{d} = m_1\mathbf{g}_1 + m_2\mathbf{g}_2 + \dots + m_M\mathbf{g}_M \tag{2.23}$$

where

$$\mathbf{m} = [m_1, m_2, \dots, m_M]^T \tag{2.24}$$

and

$$\mathbf{g}_i = [g_{1i}, g_{2i}, \dots, g_{Ni}]^T \tag{2.25}$$

is the  $i$ th column of  $\mathbf{G}$ . Also, if  $\mathbf{G}\mathbf{A} = \mathbf{B}$ , then the above can be used to infer that the first column of  $\mathbf{B}$  is a weighted sum of the columns of  $\mathbf{G}$  with the elements of the first column of  $\mathbf{A}$  as weighting factors, etc. for the other columns of  $\mathbf{B}$ . Each column of  $\mathbf{B}$  is a weighted sum of the *columns* of  $\mathbf{G}$ .

Next, consider

$$\mathbf{d}^T = [\mathbf{G}\mathbf{m}]^T \tag{2.26}$$

or

$$\begin{array}{rcc} \mathbf{d}^T & = & \mathbf{m}^T \quad \mathbf{G}^T \\ 1 \times N & & 1 \times M \quad M \times N \end{array} \tag{2.27}$$

The row vector  $\mathbf{d}^T$  is the weighted sum of the *rows* of  $\mathbf{G}^T$ , with the weighting factors again being the elements in  $\mathbf{m}$ . That is,

$$\mathbf{d}^T = m_1 \mathbf{g}_1^T + m_2 \mathbf{g}_2^T + \cdots + m_M \mathbf{g}_M^T \quad (2.28)$$

Extending this to

$$\mathbf{A}^T \mathbf{G}^T = \mathbf{B}^T \quad (2.29)$$

we have that each row of  $\mathbf{B}^T$  is a weighted sum of the *rows* of  $\mathbf{G}^T$ , with the weighting factors being the elements of the appropriate *row* of  $\mathbf{A}^T$ .

In a long string of matrix multiplications such as

$$\mathbf{ABC} = \mathbf{D} \quad (2.30)$$

each column of  $\mathbf{D}$  is a weighted sum of the *columns* of  $\mathbf{A}$ , and each row of  $\mathbf{D}$  is a weighted sum of the *rows* of  $\mathbf{C}$ .

### *Orthogonal Transformations*

An orthogonal transformation is one that leaves the length of a vector unchanged. We can only talk about the length of a vector being unchanged if the dimension of the vector is unchanged. Thus, only square matrices may represent an orthogonal transformation.

Suppose  $\mathbf{L}$  is an orthogonal transformation. Then, if

$$\mathbf{Lx} = \mathbf{y} \quad (2.31)$$

where  $\mathbf{L}$  is  $N \times N$ , and  $\mathbf{x}, \mathbf{y}$  are both  $N$ -dimensional vectors. Then

$$\mathbf{x}^T \mathbf{x} = \mathbf{y}^T \mathbf{y} \quad (2.32)$$

where Equation (2.32) represents the dot product of the vectors with themselves, which is equal to the length squared of the vector. If you have ever done coordinate transformations in the past, you have dealt with an orthogonal transformation. Orthogonal transformations rotate vectors but do not change their lengths.

*Properties of orthogonal transformations.* There are several properties of orthogonal transformations that we will wish to use.

*First*, if  $\mathbf{L}$  is an  $N \times N$  orthogonal transformation, then

$$\mathbf{L}^T \mathbf{L} = \mathbf{I}_N \quad (2.33)$$

This follows from

$$\mathbf{y}^T \mathbf{y} = [\mathbf{Lx}]^T [\mathbf{Lx}]$$

$$= \mathbf{x}^T \mathbf{L}^T \mathbf{L} \mathbf{x} \quad (2.34)$$

but  $\mathbf{y}^T \mathbf{y} = \mathbf{x}^T \mathbf{x}$  by Equation (2.32). Thus,

$$\mathbf{L}^T \mathbf{L} = \mathbf{I}_N \quad (\text{Q.E.D.}) \quad (2.35)$$

*Second*, the relationship between  $\mathbf{L}$  and its inverse is given by

$$\mathbf{L}^{-1} = \mathbf{L}^T \quad (2.36)$$

and

$$\mathbf{L} = [\mathbf{L}^T]^{-1} \quad (2.37)$$

These two follow directly from Equation (2.35) above.

*Third*, the determinant of a matrix is unchanged if it is operated upon by orthogonal transformations. Recall that the determinant of a  $3 \times 3$  matrix  $\mathbf{A}$ , for example, where  $\mathbf{A}$  is given by

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad (2.38)$$

is given by

$$\begin{aligned} \det(\mathbf{A}) &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) \\ &\quad - a_{12}(a_{21}a_{33} - a_{23}a_{31}) \\ &\quad + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \end{aligned} \quad (2.39)$$

Thus, if  $\mathbf{A}$  is an  $M \times M$  matrix, and  $\mathbf{L}$  is an orthogonal transformations, and if

$$\mathbf{A}' = (\mathbf{L})\mathbf{A}(\mathbf{L})^T \quad (2.40)$$

it follows that

$$\det(\mathbf{A}) = \det(\mathbf{A}') \quad (2.41)$$

*Fourth*, the trace of a matrix is unchanged if it is operated upon by an orthogonal transformation, where trace ( $\mathbf{A}$ ) is defined as

$$\text{trace}(\mathbf{A}) = \sum_{i=1}^M a_{ii} \quad (2.42)$$

That is, the sum of the diagonal elements of a matrix is unchanged by an orthogonal transformation. Thus,

$$\text{trace}(\mathbf{A}) = \text{trace}(\mathbf{A}') \quad (2.43)$$

### *Semiorthogonal Transformations*

Suppose that the linear operator  $\mathbf{L}$  is not square, but  $N \times M$  ( $N \neq M$ ). Then  $\mathbf{L}$  is said to be semiorthogonal if and only if

$$\mathbf{L}^T \mathbf{L} = \mathbf{I}_M, \quad \text{but } \mathbf{L} \mathbf{L}^T \neq \mathbf{I}_N, \quad N > M \quad (2.44)$$

or

$$\mathbf{L} \mathbf{L}^T = \mathbf{I}_N, \quad \text{but } \mathbf{L}^T \mathbf{L} \neq \mathbf{I}_M, \quad M > N \quad (2.45)$$

where  $\mathbf{I}_N$  and  $\mathbf{I}_M$  are the  $N \times N$  and  $M \times M$  identity matrices, respectively.

A matrix cannot be both orthogonal and semiorthogonal. Orthogonal matrices must be square, and semiorthogonal matrices cannot be square. Furthermore, if  $\mathbf{L}$  is a square  $N \times N$  matrix, and

$$\mathbf{L}^T \mathbf{L} = \mathbf{I}_N \quad (2.35)$$

then it is not possible to have

$$\mathbf{L} \mathbf{L}^T \neq \mathbf{I}_N \quad (2.46)$$

## 2.2.3 Matrices and Vector Spaces

The columns or rows of a matrix can be thought of as vectors. For example, if  $\mathbf{A}$  is an  $N \times M$  matrix, each column can be thought of as a vector in  $N$ -space because it has  $N$  entries. Conversely, each row of  $\mathbf{A}$  can be thought of as being a vector in  $M$ -space because it has  $M$  entries.

We note that for the linear system of equations given by

$$\mathbf{G} \mathbf{m} = \mathbf{d} \quad (1.13)$$

where  $\mathbf{G}$  is  $N \times M$ ,  $\mathbf{m}$  is  $M \times 1$ , and  $\mathbf{d}$  is  $N \times 1$ , that the model parameter vector  $\mathbf{m}$  lies in  $M$ -space (along with all the rows of  $\mathbf{G}$ ), while the data vector lies in  $N$ -space (along with all the columns of  $\mathbf{G}$ ). In general, we will think of the  $M \times 1$  vectors as lying in *model space*, while the  $N \times 1$  vectors lie in *data space*.

*Spanning a Space*

The notion of spanning a space is important for any discussion of the uniqueness of solutions or of the ability to fit the data. We first need to introduce definitions of linear independence and vector orthogonality.

A set of  $M$  vectors  $\mathbf{v}_i$ ,  $i = 1, \dots, M$ , in  $M$ -space (the set of all  $M$ -dimensional vectors), is said to be linearly independent if and only if

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \dots + a_M\mathbf{v}_M = 0 \quad (2.47)$$

where  $a_i$  are constants, has only the trivial solution  $a_i = 0$ ,  $i = 1, \dots, M$ .

This is equivalent to saying that an arbitrary vector  $\mathbf{s}$  in  $M$  space can be written as a linear combination of the  $\mathbf{v}_i$ ,  $i = 1, \dots, M$ . That is, one can find  $a_i$  such that for an arbitrary vector  $\mathbf{s}$

$$\mathbf{s} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \dots + a_M\mathbf{v}_M \quad (2.48)$$

Two vectors  $\mathbf{r}$  and  $\mathbf{s}$  in  $M$ -space are said to be orthogonal to each other if their dot, or inner, product with each other is zero. That is, if

$$\mathbf{r} \cdot \mathbf{s} = \|\mathbf{r}\| \|\mathbf{s}\| \cos \theta = 0 \quad (2.49)$$

where  $\theta$  is the angle between the vectors, and  $\|\mathbf{r}\|$ ,  $\|\mathbf{s}\|$  are the lengths of  $\mathbf{r}$  and  $\mathbf{s}$ , respectively.

The dot product of two vectors is also given by

$$\mathbf{r}^T \mathbf{s} = \mathbf{s}^T \mathbf{r} = \sum_{i=1}^M r_i s_i \quad (2.50)$$

$M$  space is spanned by any set of  $M$  linearly independent  $M$ -dimensional vectors.

### *Rank of a Matrix*

The number of linearly independent rows in a matrix, which is also equal to the number of linearly independent columns, is called the rank of the matrix. The rank of matrices is defined for both square and nonsquare matrices. The rank of a matrix cannot exceed the minimum of the number of rows or columns in the matrix (i.e., the rank is less than or equal to the minimum of  $N$ ,  $M$ ).

If an  $M \times M$  matrix is an orthogonal matrix, then it has rank  $M$ . The  $M$  rows are all linearly independent, as are the  $M$  columns. In fact, not only are the rows independent for an orthogonal matrix, they are orthogonal to each other. The same is true for the columns. If a matrix is semiorthogonal, then the  $M$  columns (or  $N$  rows, if  $N < M$ ) are orthogonal to each other.

We will make extensive use of matrices and linear algebra in this course, especially when we work with the generalized inverse. Next, we need to turn our attention to probability and statistics.

## 2.3 Probability and Statistics

### 2.3.1 Introduction

We need some background in *probability* and *statistics* before proceeding very far. In this review section, I will cover the material from Menke's book, using some material from other math texts to help clarify things.

Basically, what we need is a way of describing the *noise* in data and estimated model parameters. We will need the following terms: *random variable*, *probability distribution*, *mean* or *expected value*, *maximum likelihood*, *variance*, *standard deviation*, *standardized normal variables*, *covariance*, *correlation coefficients*, *Gaussian distributions*, and *confidence intervals*.

### 2.3.2 Definitions, Part 1

**Random Variable:** A function that assigns a value to the outcome of an experiment. A random variable has well-defined properties based on some distribution. It is called random because you cannot know beforehand the exact value for the outcome of the experiment. One cannot measure directly the true properties of a random variable. One can only make measurements, also called *realizations*, of a random variable, and estimate its properties. The birth weight of baby goslings is a random variable, for example.

**Probability Density Function:** The true properties of a random variable  $b$  are specified by the *probability density function*  $P(b)$ . The probability that a particular realization of  $b$  will fall between  $b$  and  $b + db$  is given by  $P(b)db$ . (Note that Menke uses  $d$  where I use  $b$ . His notation is bad when one needs to use integrals.)  $P(b)$  satisfies

$$1 = \int_{-\infty}^{+\infty} P(b) db \quad (2.51)$$

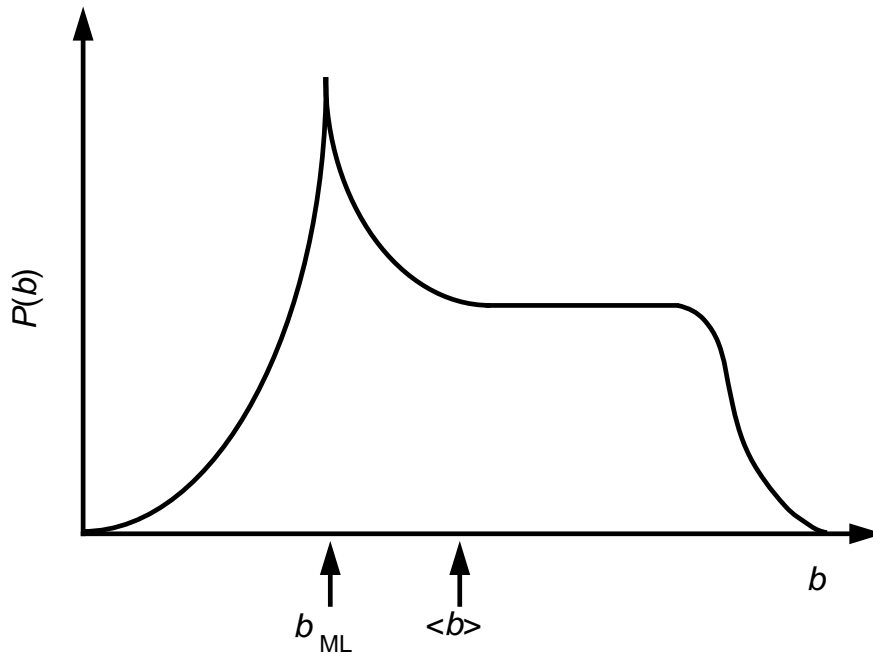
which says that the probability of  $b$  taking on some value is 1.  $P(b)$  completely describes the random variable  $b$ . It is often useful to try and find a way to summarize the properties of  $P(b)$  with a few numbers, however.

**Mean or Expected Value:** The *mean value*  $E(b)$  (also denoted  $\langle b \rangle$ ) is much like the mean of a set of numbers; that is, it is the “balancing point” of the distribution  $P(b)$  and is given by

$$E(b) = \int_{-\infty}^{+\infty} b P(b) db \quad (2.52)$$

**Maximum Likelihood:** This is the point in the probability distribution  $P(b)$  that has the highest likelihood or probability. It may or may not be close to the mean  $E(b) = \langle b \rangle$ . An important point is that for Gaussian distributions, the maximum likelihood point and the mean  $E(b) = \langle b \rangle$

are the same! The graph below (after Figure 2.3, p. 23, Menke) illustrates a case where the two are different.



The maximum likelihood point  $b_{ML}$  of the probability distribution  $P(b)$  for data  $b$  gives the most probable value of the data. In general, this value can be different from the mean datum  $\langle b \rangle$ , which is at the “balancing point” of the distribution.

**Variance:** Variance is one measure of the spread, or width, of  $P(b)$  about the mean  $E(b)$ . It is given by

$$\sigma^2 = \int_{-\infty}^{+\infty} (b - \langle b \rangle)^2 P(b) db \quad (2.53)$$

Computationally, for  $L$  experiments in which the  $k$ th experiment gives  $b_k$ , the variance is given by

$$\sigma^2 = \frac{1}{L-1} \sum_{k=1}^L (b_k - \langle b \rangle)^2 \quad (2.54)$$

**Standard Deviation:** Standard deviation is the positive square root of the variance, given by

$$\sigma = +\sqrt{\sigma^2} \quad (2.55)$$

**Covariance:** Covariance is a measure of the correlation between errors. If the errors in two observations are uncorrelated, then the covariance is zero. We need another definition before proceeding.

**Joint Density Function  $P(\mathbf{b})$ :** The probability that  $b_1$  is between  $b_1$  and  $b_1 + db_1$ , that  $b_2$  is between  $b_2$  and  $b_2 + db_2$ , etc. If the data are independent, then

$$P(\mathbf{b}) = P(b_1) P(b_2) \cdots P(b_n) \quad (2.56)$$

If the data are correlated, then  $P(\mathbf{b})$  will have some more complicated form. Then, the covariance between  $b_1$  and  $b_2$  is defined as

$$\text{cov}(b_1, b_2) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (b_1 - \langle b_1 \rangle)(b_2 - \langle b_2 \rangle) P(\mathbf{b}) db_1 db_2 \cdots db_n \quad (2.57)$$

In the event that the data are independent, this reduces to

$$\begin{aligned} \text{cov}(b_1, b_2) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (b_1 - \langle b_1 \rangle)(b_2 - \langle b_2 \rangle) P(b_1) P(b_2) db_1 db_2 \\ &= 0 \end{aligned} \quad (2.58)$$

The reason is that for any value of  $(b_1 - \langle b_1 \rangle)$ ,  $(b_2 - \langle b_2 \rangle)$  is as likely to be positive as negative, i.e., the sum will average to zero. The matrix  $[\text{cov } \mathbf{b}]$  contains all of the covariances defined using Equation (2.57) in an  $N \times N$  matrix. Note also that the covariance of  $b_i$  with itself is just the variance of  $b_i$ .

In practical terms, if one has an  $N$ -dimensional data vector  $\mathbf{b}$  that has been measured  $L$  times, then the  $ij$ th term in  $[\text{cov } \mathbf{b}]$ , denoted  $[\text{cov } \mathbf{b}]_{ij}$ , is defined as

$$[\text{cov } \mathbf{b}]_{ij} = \frac{1}{L-1} \sum_{k=1}^L (b_i^k - \bar{b}_i)(b_j^k - \bar{b}_j) \quad (2.59)$$

where  $b_i^k$  is the value of the  $i$ th datum in  $\mathbf{b}$  on the  $k$ th measurement of the data vector,  $\bar{b}_i$  is the mean or average value of  $b_i$  for all  $L$  measurements (also commonly written  $\langle b_i \rangle$ ), and the  $L - 1$  term results from sampling theory.

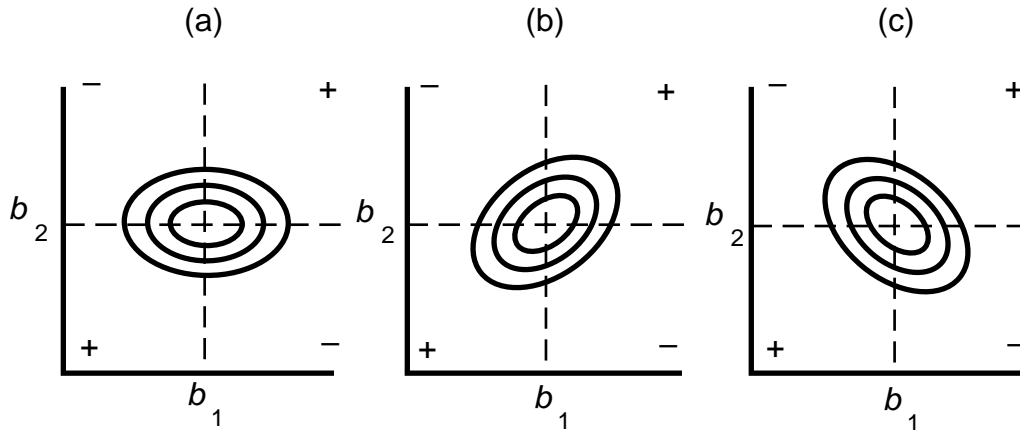
**Correlation Coefficients:** This is a normalized measure of the degree of correlation of errors. It takes on values between  $-1$  and  $1$ , with a value of  $0$  implying no correlation.

The correlation coefficient matrix  $[\text{cor } \mathbf{b}]$  is defined as

$$[\text{cor } \mathbf{b}]_{ij} = \frac{[\text{cov } \mathbf{b}]_{ij}}{\sigma_i \sigma_j} \quad (2.60)$$

where  $[\text{cov } \mathbf{b}]_{ij}$  is the covariance matrix defined term by term as above for  $\text{cov}[b_1, b_2]$ , and  $\sigma_i, \sigma_j$  are the standard deviations for the  $i$ th and  $j$ th observations, respectively. The diagonal terms of  $[\text{cor } \mathbf{b}]_{ij}$  are equal to  $1$ , since each observation is perfectly correlated with itself.

The figure below (after Figure 2.8, page 26, Menke) shows three different cases of degree of correlation for two observations  $b_1$  and  $b_2$ .



Contour plots of  $P(b_1, b_2)$  when the data are (a) uncorrelated, (b) positively correlated, (c) negatively correlated. The dashed lines indicate the four quadrants of alternating sign used to determine correlation.

### 2.3.3 Some Comments on Applications to Inverse Theory

Some comments are now in order about the nature of the estimated model parameters. We will always assume that the noise in the observations can be described as random variables. Whatever inverse we create will map errors in the data into errors in the estimated model parameters. Thus, the estimated model parameters are themselves random variables. This is true even though the true model parameters may not be random variables. If the distribution of noise for the data is known, then in principle the distribution for the estimated model parameters can be found by “mapping” through the inverse operator.

This is often very difficult, but one particular case turns out to have a rather simple form. We will see where this form comes from when we get to the subject of generalized inverses. For now, consider the following as magic.

If the transformation between data  $\mathbf{b}$  and model parameters  $\mathbf{m}$  is of the form

$$\mathbf{m} = \mathbf{M}\mathbf{b} + \mathbf{v} \tag{2.61}$$

where  $\mathbf{M}$  is any arbitrary matrix and  $\mathbf{v}$  is any arbitrary vector, then

$$\langle \mathbf{m} \rangle = \mathbf{M}\langle \mathbf{b} \rangle + \mathbf{v} \tag{2.62}$$

and

$$[\text{cov } \mathbf{m}] = \mathbf{M} [\text{cov } \mathbf{b}] \mathbf{M}^T \tag{2.63}$$

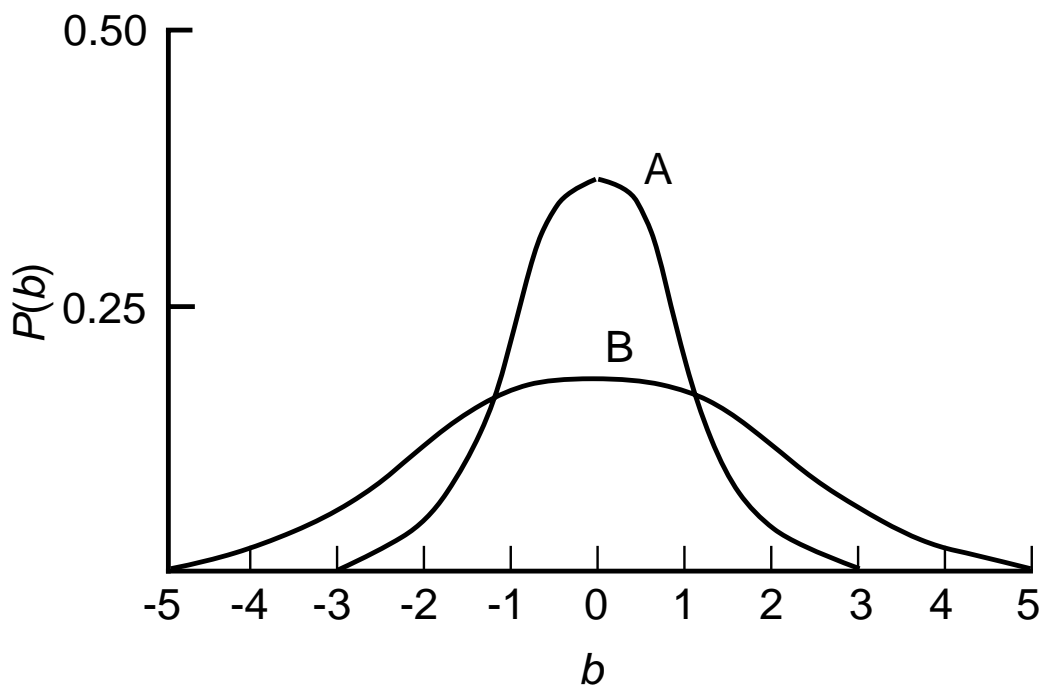
### 2.3.4 Definitions, Part 2

**Gaussian Distribution:** This is a particular probability distribution given by

$$P(b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(b - \langle b \rangle)^2}{2\sigma^2}\right] \quad (2.64)$$

The figure below (after Figure 2.10, page 29, Menke) shows the familiar bell-shaped curve. It has the following properties:

$$\text{Mean} = E(b) = \langle b \rangle \quad \text{and} \quad \text{Variance} = \sigma^2$$



Gaussian distribution with zero mean and  $\sigma = 1$  for curve A, and  $\sigma = 2$  for curve B.

Many distributions can be approximated fairly accurately (especially away from the tails) by the Gaussian distribution. It is also very important because it is the limiting distribution for the sum of random variables. This is often just what one assumes for noise in the data.

One also needs a way to represent the joint probability introduced earlier for a set of random variables each of which has a Gaussian distribution. The joint probability density function for a vector  $\mathbf{b}$  of observations that all have Gaussian distributions is chosen to be [see Equation (2.10) of Menke, page 30]

$$P(\mathbf{b}) = \frac{(\det[\text{cov } \mathbf{b}])^{-1/2}}{(2\pi)^{N/2}} \exp\left\{-\frac{1}{2}[\mathbf{b} - \langle \mathbf{b} \rangle]^T [\text{cov } \mathbf{b}]^{-1} [\mathbf{b} - \langle \mathbf{b} \rangle]\right\} \quad (2.65)$$

which reduces to the previous case in Equation (2.64) for  $N = 1$  and  $\text{var}(b_1) = \sigma^2$ . In statistics books, Equation (2.65) is often given as

$$P(\mathbf{b}) = (2\pi)^{-N/2} |\Sigma_{\mathbf{b}}|^{-1/2} \exp\{-2[\mathbf{b} - \boldsymbol{\mu}_{\mathbf{b}}]^T \Sigma^{-1}[\mathbf{b} - \boldsymbol{\mu}_{\mathbf{b}}]\}$$

With this background, it makes sense (statistically, at least) to replace the original relationship:

$$\mathbf{b} = \mathbf{G}\mathbf{m} \tag{1.13}$$

with

$$\langle \mathbf{b} \rangle = \mathbf{G}\mathbf{m} \tag{2.66}$$

The reason is that one cannot expect that there is an  $\mathbf{m}$  that should exactly predict any particular realization of  $\mathbf{b}$  when  $\mathbf{b}$  is in fact a random variable.

Then the joint probability is given by

$$P(\mathbf{b}) = \frac{(\det[\text{cov } \mathbf{b}])^{-1/2}}{(2\pi)^{N/2}} \exp\left\{-\frac{1}{2}[\mathbf{b} - \mathbf{G}\mathbf{m}]^T [\text{cov } \mathbf{b}]^{-1}[\mathbf{b} - \mathbf{G}\mathbf{m}]\right\} \tag{2.67}$$

What one then does is seek an  $\mathbf{m}$  that maximizes the probability that the predicted data are in fact close to the observed data. This is the basis of the *maximum likelihood* or probabilistic approach to inverse theory.

**Standardized Normal Variables:** It is possible to standardize random variables by subtracting their mean and dividing by the standard deviation.

If the random variable had a Gaussian (i.e., normal) distribution, then so does the standardized random variable. Now, however, the standardized normal variables have zero mean and standard deviation equal to one. Random variables can be standardized by the following transformation:

$$s = \frac{\mathbf{m} - \langle \mathbf{m} \rangle}{\sigma} \tag{2.68}$$

where you will often see  $\mathbf{z}$  replacing  $s$  in statistics books.

We will see, when all is said and done, that most inverses represent a transformation to standardized variables, followed by a “simple” inverse analysis, and then a transformation back for the final solution.

**Chi-Squared (Goodness of Fit) Test:** A statistical test to see whether a particular observed distribution is likely to have been drawn from a population having some known form.

The application we will make of the chi-squared test is to test whether the noise in a particular problem is likely to have a Gaussian distribution. This is not the kind of question one can answer with certainty, so one must talk in terms of probability or likelihood. For example, in the chi-squared test, one typically says things like there is only a 5% chance that this sample distribution does not follow a Gaussian distribution.

As applied to testing whether a given distribution is likely to have come from a Gaussian population, the procedure is as follows: One sets up an arbitrary number of bins and compares the number of observations that fall into each bin with the number expected from a Gaussian distribution having the same mean and variance as the observed data. One quantifies the departure between the two distributions, called the chi-squared value and denoted  $\chi^2$ , as

$$\chi^2 = \sum_{i=1}^k \frac{[(\# \text{obs in bin } i) - (\# \text{expected in bin } i)]^2}{[\# \text{expected in bin } i]} \quad (2.69)$$

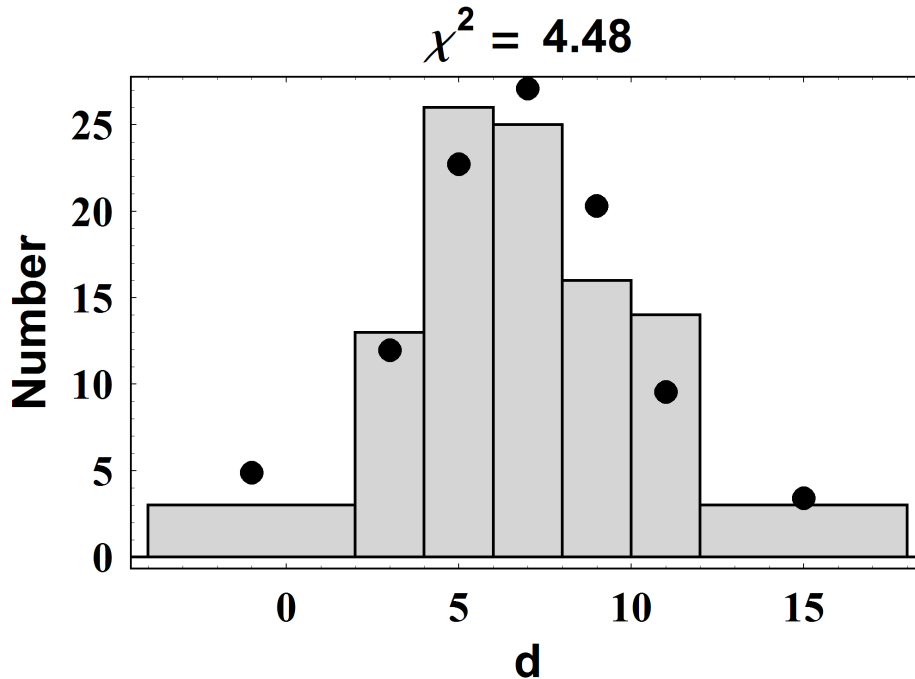
where the sum is over the number of bins,  $k$ . Next, the number of degrees of freedom for the problem must be considered. For this problem, the number of degrees is equal to the number of bins minus three. The reason you subtract three is as follows: You subtract 1 because if an observation does not fall into any subset of  $k - 1$  bins, you know it falls in the one bin left over. You are not free to put it anywhere else. The other two come from the fact that you have assumed that the mean and standard deviation of the observed data set are the mean and standard deviations for the theoretical Gaussian distribution.

With this information in hand, one uses standard chi-squared test tables from statistics books and determines whether such a departure would occur randomly more often than, say, 5% of the time. Officially, the null hypothesis is that the sample was drawn from a Gaussian distribution. If the observed value for  $\chi^2$  is greater than  $\chi^2_{\alpha}$ , called the critical  $\chi^2$  value for the  $\alpha$  significance level, then the null hypothesis is rejected at the  $\alpha$  significance level. Commonly,  $\alpha = 0.05$  is used for this test, although  $\alpha = 0.01$  is also used. The  $\alpha$  significance level is equivalent to the  $100 \cdot (1 - \alpha)\%$  confidence level (i.e.,  $\alpha = 0.05$  corresponds to the 95% confidence level).

Consider the following example, where the underlying Gaussian distribution from which all data samples  $d$  are drawn has a mean of 7 and a variance of 10. Seven bins are set up with edges at  $-4, 2, 4, 6, 8, 10, 12,$  and  $18$ , respectively. Bin widths are not prescribed for the chi-squared test, but ideally are chosen so there are about an equal number of occurrences expected in each bin. Also, one rule of thumb is to only include bins having at least five expected occurrences. I have not followed the “about equal number expected in each bin” suggestion because I want to be able to visually compare a histogram with an underlying Gaussian shape. However, I have chosen wider bins at the edges in these test cases to capture more occurrences at the edges of the distribution.

Suppose our experiment with 100 observations yields a sample mean of 6.76 and a sample variance of 8.27, and 3, 13, 26, 25, 16, 14, and 3 observations, respectively, in the bins from left to right. Using standard formulas for a Gaussian distribution with a mean of 6.76 and a variance of 8.27, the number expected in each bin is 4.90, 11.98, 22.73, 27.10, 20.31, 9.56, and 3.41, respectively. The calculated  $\chi^2$ , using Equation (2.69), is 4.48. For seven bins, the DOFs

for the test is 4, and  $\chi^2_{\alpha} = 9.49$  for  $\alpha = 0.05$ . Thus, in this case, the null hypothesis would be accepted. That is, we would accept that this sample was drawn from a Gaussian distribution with a mean of 6.76 and a variance of 8.27 at the  $\alpha = 0.05$  significance level (95% confidence level). The distribution is shown below, with a filled circle in each histogram at the number expected in that bin.



It is important to note that this distribution does not look exactly like a Gaussian distribution, but still passes the  $\chi^2$  test. A simple, non-chi-square analogy may help better understand the reasoning behind the chi-square test. Consider tossing a true coin 10 times. The most likely outcome is 5 heads and 5 tails. Would you reject a null hypothesis that the coin is a true coin if you got 6 heads and 4 tails in your one experiment of tossing the coin ten times? Intuitively, you probably would not reject the null hypothesis in this case, because 6 heads and 4 tails is “not that unlikely” for a true coin.

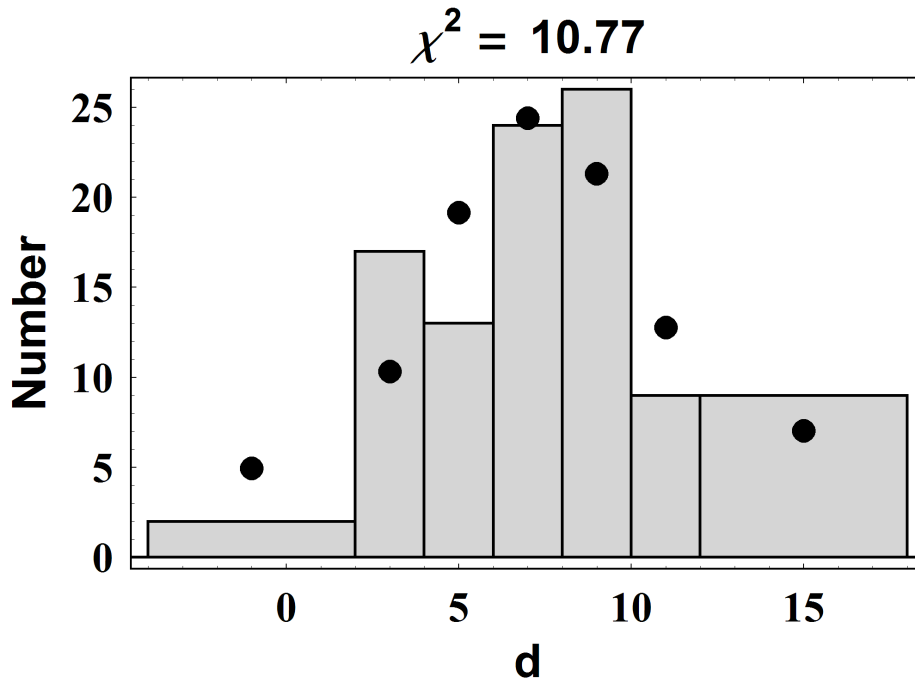
In order to make an informed decision, as we try to do with the chi-square test, you would need to quantify how likely, or unlikely, a particular outcome is before accepting or rejecting the null hypothesis that it is a true coin. For a true coin, 5 heads and 5 tails has a probability of 0.246 (that is, on average, it happens 24.6% of the time), while the probability of 6 heads and 4 tails is 0.205, 7 heads and 3 tails is 0.117, and 8 heads and 2 tails is 0.044, respectively. A distribution of 7 heads and 3 tails does not “look” like 5 heads and 5 tails, but occurs more than 10% of the time with a true coin.

Hence, by analogy, it is not “too unlikely” and you would probably not reject the null hypothesis that the coin is a true coin just because you tossed 7 heads and 3 tails in one experiment. Ten heads and no tails only occurs, on average, one time in 1024 experiments (or about 0.098% of the time). If you got 10 heads and 0 tails, you’d probably reject the null hypothesis that you are tossing a true coin because the outcome is very unlikely. Eight heads and two tails occurs 4.4% of the time, on average. You might also reject the null hypothesis in this

case, but you would do so with less confidence, or at a lower significance level. In both cases, however, your conclusion will be wrong occasionally just due to random variations. You accept the possibility that you will be wrong rejecting the null hypothesis 4.4% of the time in this case, even if the coin is true.

The same is true with the chi-square test. That is, at the  $\alpha = 0.05$  significance level (95% confidence level), with  $\chi^2$  greater than  $\chi^2_{\alpha}$  you reject the null hypothesis, even though you recognize that you will reject the null hypothesis incorrectly about 5% of the time in the presence of random variations. Note that this analogy is a simple one in the sense that it is entirely possible to actually do a chi-square test on this coin toss example. Each time you toss the coin ten times you get one outcome:  $x$  heads and  $(10 - x)$  tails. This falls into the “ $x$  heads and  $(10 - x)$  tails” bin. If you repeat this many times you get a distribution across all bins from “0 heads and 10 tails” to “10 heads and 0 tails.” Then you would calculate the number expected in each bin and use Equation (2.69) to calculate a chi-square value to compare with the critical value at the  $\alpha$  significance level.

Now let us return to another example of the chi-square test where we reject the null hypothesis. Consider a case where the observed number in each of the seven bins defined above is now 2, 17, 13, 24, 26, 9, and 9, respectively, and the observed distribution has a mean of 7.28 and variance of 10.28. The expected number in each bin, for the observed mean and variance, is 4.95, 10.32, 19.16, 24.40, 21.32, 12.78, and 7.02, respectively. The calculated  $\chi^2$  is now 10.77, and the null hypothesis would be rejected at the  $\alpha = 0.05$  significance level (95% confidence level). That is, we would reject that this sample was drawn from a Gaussian distribution with a mean of 7.28 and variance of 10.28 at this significance level. The distribution is shown on the next page, again with a filled circle in each histogram at the number expected in that bin.



**Confidence Intervals:** One says, for example, with 98% confidence that the true mean of a random variable lies between two values. This is based on knowing the probability distribution

for the random variable, of course, and can be very difficult, especially for complicated distributions that include nonzero correlation coefficients. However, for Gaussian distributions, these are well known and can be found in any standard statistics book. For example, Gaussian distributions have 68% and 95% confidence intervals of approximately  $\pm 1\sigma$  and  $\pm 2\sigma$ , respectively.

***T* and *F* Tests:** These two statistical tests are commonly used to determine whether the properties of two samples are consistent with the samples coming from the same population.

The *F* test in particular can be used to test the improvement in the fit between predicted and observed data when one adds a degree of freedom in the inversion. One expects to fit the data better by adding more model parameters, so the relevant question is whether the improvement is significant.

As applied to the test of improvement in fit between case 1 and case 2, where case 2 uses more model parameters to describe the same data set, the *F* ratio is given by

$$F = \frac{(E_1 - E_2)/(DOF_1 - DOF_2)}{(E_2 / DOF_2)} \quad (2.70)$$

where *E* is the residual sum of squares and *DOF* is the number of degrees of freedom for each case.

If *F* is large, one accepts that the second case with more model parameters provides a significantly better fit to the data. The calculated *F* is compared to published tables with *DOF*<sub>1</sub> – *DOF*<sub>2</sub> and *DOF*<sub>2</sub> degrees of freedom at a specified confidence level. (Reference: T. M. Hearn, *P<sub>n</sub>* travel times in Southern California, *J. Geophys. Res.*, 89, 1843–1855, 1984.)

---

The next section will deal with solving inverse problems based on length measures. This will include the classic least squares approach.