Chapter 3: Calculus Tools

Prerequisite: Chapter 1

3.1 *Logarithms and Exponents*

By definition, the log function to the base b is the function such that $c = \log_b a$ if $b^c = a$. It is a very useful function in statistical reasoning, since it takes multiplication into addition as we will see in Equation (3.1). We generally use the notation *log* to imply a base of 10, i. e. $\log a = \log_{10} a$ and we use the notation *ln* to imply a base of Euler's e (2.7182812…), that is $\ln a = \log_e a$. Some rules of logarithms follow:

$$\ln ab = \ln a + \ln b \tag{3.1}$$

$$\ln \frac{a}{b} = \ln a - \ln b \tag{3.2}$$

$$\ln a^b = b \ln a \tag{3.3}$$

$$\ln e^a = a \tag{3.4}$$

$$\ln e = 1 \tag{3.5}$$

$$\ln 1 = 0 \tag{3.6}$$

As for exponents, we have the following rules:

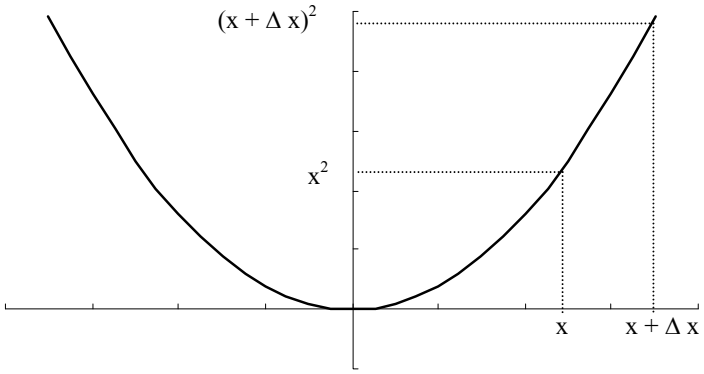
$$a^b \cdot a^c = a^{b+c} \tag{3.7}$$

$$a^{1/2} = \sqrt{a} \tag{3.8}$$

From a purely typographical point of view, it is sometimes more convenient to use the notation $\exp(a) = e^a$.

3.2 *A Review of Scalar Calculus*

Consider the problem of calculating the slope of $f(x) = x^2$. Unlike the equation for a line, the slope of the f(x) function changes depending on the value of x. However, at a small enough segment of the function, fitting a straight line would be reasonable. A picture of the situation is given below:



The slope is composed of the amount of change in the y axis (the rise) divided by the change in the x axis (the run). The fraction looks like

$$\text{slope} = \frac{(x+\Delta x)^2 - x^2}{(x+\Delta x) - x}$$

$$= \frac{x^2 + 2x\cdot\Delta x + (\Delta x)^2 - x^2}{\Delta x}$$

$$= 2x + \Delta x.$$

As we reduce Δx smaller and smaller, making a closer approximation to the slope, it converges on the value 2x. The derivative is the slope of a function at a point. There are two notations in common use. Thus we could write $dx^2/dx = 2x$ or $f'(x) = 2x$. In this book we will generally stick to the first way of writing the derivative.

More generally, for a function consisting of the power of a variable,

$$\frac{dx^m}{dx} = m\cdot x^{m-1}. \tag{3.9}$$

For the function f(x) = c where c is a constant, we would have

$$d(c)/dx = 0 \tag{3.10}$$

and for f(x) = cx,

$$d(cx)/dx = c. \tag{3.11}$$

The derivative of a sum is equal to the sum of the derivatives as we now see:

$$\frac{d[f(x)+g(x)]}{dx} = \frac{df(x)}{dx} + \frac{dg(x)}{dx}. \tag{3.12}$$

The exponential function to the base e has the interesting property that

$$\frac{de^x}{dx} = e^x \tag{3.13}$$

And we finish up this review by noting that for compound functions, such as g[f(x)], we can employ the *chain rule* which states that

$$\frac{dg[f(x)]}{dx} = \frac{dg[f(x)]}{df(x)} \cdot \frac{df(x)}{dx}. \tag{3.14}$$

Now, taking the chain rule into account we can state

$$\frac{de^{f(x)}}{dx} = e^{f(x)} \cdot \frac{df(x)}{dx}. \tag{3.15}$$

3.3 *The Scalar Function of a Vector*

We can now define the derivative of a function with respect to a whole vector of "independent" variables, $\partial f(\mathbf{x}')/\partial \mathbf{x}'$. Note that the function of the vector, f(**x**'), is a scalar. To begin, we will start with the constant function, that is, f(**x**') = c where c is a constant (scalar). The derivative of this function with respect to the row vector **x**' is itself a row vector with the same order as **x**'. That is because we need a derivative of the function with respect to each element of the vector. This vector derivative is called a *partial derivative* which means that as we take the derivative of the function with respect x_i, each of the other elements of **x** are treated as constants.

$$\begin{aligned}\frac{\partial c}{\partial \mathbf{x}'} &= \left[\frac{\partial c}{\partial x_1} \quad \frac{\partial c}{\partial x_2} \quad \cdots \quad \frac{\partial c}{\partial x_m}\right] \\ &= [0 \quad 0 \quad \cdots \quad 0].\end{aligned} \tag{3.16}$$

The derivative of the function with respect to x_i is 0, and i runs from 1 to m. Thus a vector derivative is created. For the linear combination **a'x** we have

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}'} = \left[\frac{\partial \mathbf{a}'\mathbf{x}}{\partial x_1} \quad \frac{\partial \mathbf{a}'\mathbf{x}}{\partial x_2} \quad \cdots \quad \frac{\partial \mathbf{a}'\mathbf{x}}{\partial x_m}\right]$$

$$= \left[\frac{\partial}{\partial x_1}[a_1x_1 + a_2x_2 + \cdots + a_mx_m] \quad \frac{\partial}{\partial x_2}[a_1x_1 + a_2x_2 + \cdots + a_mx_m] \quad \cdots \quad \frac{\partial}{\partial x_m}[a_1x_1 + a_2x_2 + \cdots + a_mx_m]\right]$$

$$= [a_1 \quad a_2 \quad \cdots \quad a_m] = \mathbf{a}'. \tag{3.17}$$

Another important result is the derivative of a quadratic form [Equation (1.20)]. In the equation below, we assume that **A** is a symmetric m · m matrix so that

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} = 2\mathbf{x}'\mathbf{A}' \tag{3.18}$$

with $\mathbf{A}' = \mathbf{A}$.

We now state the rule that the derivative of the transpose is equal to the transpose of the derivative, that is

$$\frac{\partial \mathrm{f}}{\partial \mathbf{x}'} = \left[\frac{\partial \mathrm{f}}{\partial \mathbf{x}}\right]' \text{ and}$$
$$\frac{\partial \mathrm{f}}{\partial \mathbf{x}} = \left[\frac{\partial \mathrm{f}}{\partial \mathbf{x}'}\right]' \tag{3.19}$$

From time to time we will need to use the *second order derivative* of a scalar function. It may be the case that the $\partial f/\partial x_i$ changes as a function of x_j, for example. The slope of the $\partial f/\partial x_i$ with respect to x_j, in other words the derivative of the derivative, is written as

$$\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right) = \frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

There are many uses of this second order derivative including nonlinear estimation [Section (3.9)], Maximum Likelihood parameter estimation [Section (3.10)], as well as determining whether, when the first order derivative is 0, we are at a maximum or minimum.

3.4 *Derivative of Multiple Functions with Respect to a Vector*

Suppose we have the linear system,

$$_{\mathrm{n}}\mathbf{y}_{1} = {}_{\mathrm{n}}\mathbf{A}_{\mathrm{m}\,\mathrm{m}}\,\mathbf{x}_{1}\,.$$

Now

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} = \frac{\partial}{\partial \mathbf{x}'}\begin{bmatrix}\mathbf{a}'_{1\cdot}\\ \mathbf{a}'_{2\cdot}\\ \cdots\\ \mathbf{a}'_{\mathrm{n}\cdot}\end{bmatrix}\mathbf{x}$$

$$= \begin{bmatrix}\frac{\partial \mathbf{a}'_{1\cdot}}{\partial \mathbf{x}'}\\ \frac{\partial \mathbf{a}'_{2\cdot}}{\partial \mathbf{x}'}\\ \cdots\\ \frac{\partial \mathbf{a}'_{\mathrm{n}\cdot}}{\partial \mathbf{x}'}\end{bmatrix} = \begin{bmatrix}\mathbf{a}'_{1\cdot}\\ \mathbf{a}'_{2\cdot}\\ \cdots\\ \mathbf{a}'_{\mathrm{n}\cdot}\end{bmatrix} = \mathbf{A}$$

To summarize,

$$\frac{\partial_n \mathbf{y}_1}{\partial_1 \mathbf{x}'_m} = \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} = {}_n\mathbf{A}_m \tag{3.20}$$

with each of the n rows of $\partial\mathbf{y}/\partial\mathbf{x}'$ a different function of $\mathbf{x}'$, $y_1, y_2, \ldots, y_n$ and each of the m columns of $\partial\mathbf{y}/\partial\mathbf{x}'$ referring to a different independent variable: $x_1, x_2, \ldots, x_m$. In other words, element i, j of $\partial\mathbf{y}/\partial\mathbf{x}'$ is of $\partial y_i/\partial x_j = a_{ij}$.

Of course given Equation (3.19),

$$\frac{\partial \mathbf{y}'}{\partial \mathbf{x}} = \left[\frac{\partial \mathbf{y}}{\partial \mathbf{x}'}\right]' = {}_m\mathbf{A}'_n \, .$$

3.5 *Eigen Structure for Symmetric Matrices*

Consider the p by 1 random vector $\mathbf{y}$, consisting of p observations taken on a randomly chosen case. The covariance matrix $\mathbf{S}$, which is the covariance matrix for p variables [that is, $V(\mathbf{y}) = \mathbf{S}$], is a symmetric matrix. I wish to create a linear combination

$$u = \mathbf{x}'\mathbf{y}, \tag{3.21}$$

such that q = V(u) is maximized. In this way I can replace the p elements of $\mathbf{y}$ with a single number that behaves as much as possible like the original p values. The problem can be written as

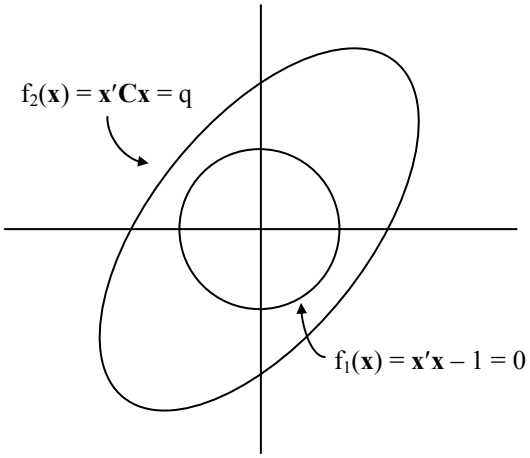
$$\frac{\text{Max } q = \mathbf{x}'\mathbf{S}\mathbf{x}}{\mathbf{x}} \, . \tag{3.22}$$

The notation here can be read, "Max q over all values of $\mathbf{x}$." One easy way to do this would be to pick $\mathbf{x} = [\infty \; \infty \; \cdots \; \infty]'$ but this would be completely uninteresting. Instead we will normalize $\mathbf{x}$, or constrain it so that we do not fall into a solution with a series of infinities. The reasoning behind how we maximize a function under constraints was introduced into mathematics by Lagrange. We can arbitrarily fix

$$\sum_j^p x_j^2 = \mathbf{x}'\mathbf{x} = 1 \text{ or set}$$

$$\mathbf{x}'\mathbf{x} - 1 = 0 \, . \tag{3.23}$$

This will allow us to focus on the pattern in the $\mathbf{x}$ vector that allows us to extract the maximum variance from $\mathbf{S}$. Geometrically, we can represent the situation as in the graph below:



Rather than trying to maximize $f_2(\mathbf{x})$, we will maximize $f_2(\mathbf{x})$ subject to $f_1(\mathbf{x})$. This is equivalent to maximizing $f_2(\mathbf{x})$ - $f_1(\mathbf{x})$, or finding the *principal axis* of the ellipse in the figure. The problem can now be written as

$$\underset{\mathbf{x}}{\text{Max}}\left[f_2(\mathbf{x}) - \lambda f_1(\mathbf{x})\right] = \underset{\mathbf{x}}{\text{Max}}\left[\mathbf{x}'\mathbf{S}\mathbf{x} - \lambda(\mathbf{x}'\mathbf{x} - 1)\right] \tag{3.24}$$

Note the sudden and mysterious appearance of the scalar λ! This creature is called a *Lagrange multiplier*. But where did it come from? Indeed. In defense of this equation, note that $f_2(\mathbf{x}) = \mathbf{x}'\mathbf{x}$ - 1 = 0. The scalar λ does not change the equation one iota, or better; one lambda. The function $f_2(\mathbf{x})$, as well as λ, are doomed to vanish. In short, λ is a mathematical throw-away. Using the rule for the derivative of a quadratic form [Equation (3.18)], along with some help from Equation (3.19), we see that

$$\frac{\partial \mathbf{x}'\mathbf{S}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{S}\mathbf{x}$$

and that

$$\frac{\partial \lambda(\mathbf{x}'\mathbf{x} - 1)}{\partial \mathbf{x}} = 2\lambda\mathbf{I}\mathbf{x}\,. \tag{3.25}$$

In that case, to maximize (3.24) we set

$$\frac{\partial}{\partial \mathbf{x}}\left[\mathbf{x}'\mathbf{S}\mathbf{x} - \lambda(\mathbf{x}'\mathbf{x} - 1)\right] = \mathbf{0}$$

$$2\mathbf{S}\mathbf{x} - 2\lambda\mathbf{I}\mathbf{x} = \mathbf{0}\,. \tag{3.26}$$

We can simplify further as below,

$$\mathbf{S}\mathbf{x} = \lambda\mathbf{x}, \tag{3.27}$$

where λ is now "acting like" **S**. Putting λ in Equation (3.24) is certainly legal since $\mathbf{x}'\mathbf{x}$ - 1 will be zero anyway. But what is it doing still hanging around in Equation (3.27)? We promised it would

go away, didn't we? What is λ anyway? Before we can answer this we need to return to Equation (3.27), were we had

$$\mathbf{Sx} = \lambda\mathbf{x}$$

which when premultiplied by $\mathbf{x}'$ leads to

$$\mathbf{x}'\mathbf{Sx} = \mathbf{x}'\lambda\mathbf{x}.$$

By the rules of scalar multiplication [in particular Equation (1.28)], and by the fact that $\mathbf{x}'\mathbf{x} = 1$ we have

$$\mathbf{x}'\lambda\mathbf{x} = \mathbf{x}'\mathbf{x}\lambda = \lambda$$

so that we can conclude

$$\mathbf{x}'\mathbf{Sx} = \lambda. \tag{3.28}$$

At this point the reader will recognize the formula for the variance of a linear combination, Equation 4.9. The value λ is called an *eigenvalue* of the matrix **S**. It is the maximum value, q, of the variance of $u = \mathbf{x}'\mathbf{y}$ which was our original motivation for this problem way back in Equation (3.21). The vector **x** chosen to maximize this variance is called an *eigenvector* of **S**.

3.6 *A Small Example Calculating the Eigenvalue and Eigenvector*

We will now return to Equation (3.26), which although it looked like

$$2\mathbf{Sx} - 2\lambda\mathbf{Ix} = \mathbf{0},$$

we can multiply by 1/2 to create

$$\mathbf{Sx} - \lambda\mathbf{Ix} = \mathbf{0} \quad \text{or}$$

$$[\mathbf{S} - \lambda\mathbf{I}]\,\mathbf{x} = \mathbf{0}. \tag{3.29}$$

Equation (3.29) can be solved trivially, as

$$[\mathbf{S} - \lambda\mathbf{I}]^{-1}[\mathbf{S} - \lambda\mathbf{I}]\,\mathbf{x} = [\mathbf{S} - \lambda\mathbf{I}]^{-1}\mathbf{0}$$

$$\mathbf{x} = \mathbf{0},$$

but such a solution would not be useful at all to us and in fact would not give us what we are looking for, namely, the linear combination $u = \mathbf{x}'\mathbf{y}$ such that $V(u) = \mathbf{x}'\mathbf{Sx}$ is as large as possible. To avoid falling into this trivial solution we must somehow pick λ such that

$$|\mathbf{S} - \lambda\mathbf{I}| = 0$$

which in turn implies that $[\mathbf{S} - \lambda\mathbf{I}]^{-1}$ does not exist (see Section 1.8). If $[\mathbf{S} - \lambda\mathbf{I}]^{-1}$ does not exist, we are not stuck with $\mathbf{x} = \mathbf{0}$, the trivial solution. Below, we can see how this works with a 2×2 example, lets say

$$\mathbf{S} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

In that case we have

$$|\mathbf{S} - \lambda\mathbf{I}| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0.$$

Recalling the equation for the determinant of a 2 × 2 matrix [from the denominator of Equation (1.38)], we have

$$(2 - \lambda)^2 - 1^2 = 0$$

which as a quadratic equation has two roots, i. e.

$$4 - 2\lambda - 2\lambda + \lambda^2 - 1 = 0$$

$$\lambda^2 - 4\lambda + 3 = 0$$

$$(\lambda - 3)(\lambda - 1) = 0$$

where the roots are $\lambda_1 = 3$ and $\lambda_2 = 1$. The first eigenvalue represents the maximum variance while the second represents the maximum variance that can be found after the first linear combination has been extracted. It is also true that the last eigenvalue represents the minimum amount of variance that can be extracted by a linear combination. We can now substitute λ_1 back into Equation (3.29) in order to solve for the first eigenvector. Calling this first eigenvector $\mathbf{x}_{\cdot 1}$, we have

$$[\mathbf{S} - \lambda\mathbf{I}]\,\mathbf{x}_{\cdot 1} = \mathbf{0}$$

$$\begin{bmatrix} 2-3 & 1 \\ 1 & 2-3 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

so that $-x_{11} + x_{21} = 0$ and $x_{11} - x_{21} = 0$. It is obvious then that $x_{11} = x_{21}$. Taken together with the restriction that $\mathbf{x}'\mathbf{x} = 1$ that we imposed in Equation (3.23), we have

$$\mathrm{x}_{\cdot 1} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

3.7 *Some Properties of Eigenstructure*

Before proceeding, it will be useful to take each of the eigenvectors, $\mathbf{x}_{\cdot 1}$, $\mathbf{x}_{\cdot 2}$, $\cdots$, $\mathbf{x}_{\cdot p}$, and place them as columns into the matrix $\mathbf{X}$. We also take the eigenvalues, λ_1, λ_2, $\cdots$, λ_p and put them on the diagonal of the matrix $\mathbf{L}$. The eigenvalues in $\mathbf{L}$ summarize a variety of properties of the original matrix $\mathbf{S}$. For example:

$$\mathrm{Tr}(\mathbf{S}) = \sum_{i}^{p} \lambda_i = \mathrm{tr}(\mathbf{\Lambda}) \tag{3.30}$$

$$|\mathbf{S}| = \prod_{i}^{p} \lambda_i \tag{3.31}$$

The *rank of a square matrix* $\mathbf{S}$ is given by the number of eigenvalues > 0. In other words, the rank of a square matrix is given by the number of non-null eigenvectors. We say that a square matrix is of *full rank* if one cannot pick a non-null vector $\mathbf{x}$ such that $\mathbf{x}'\mathbf{S}\mathbf{x} = 0$. We can see then from Equation (3.31) that if no eigenvalue is zero, the determinant, $|\mathbf{S}|$, will be non-zero and it will be possible to find $\mathbf{S}^{-1}$.

For each eigenvector-eigenvalue combination i, we have

$$\mathbf{S}\mathbf{x}_{\cdot i} = \mathbf{x}_{\cdot i}\lambda_i$$

so that if we premultiply by $\mathbf{x}'_{\cdot j}$ we have

$$\mathbf{x}'_{\cdot j}\mathbf{S}\mathbf{x} = \mathbf{x}'_{\cdot j}\mathbf{x}_{\cdot i}\boldsymbol{\lambda}_i \,.$$

Making the same argument for the eigenvalue and eigenvector j, we have

$$\mathbf{S}\mathbf{x}_{\cdot j} = \mathbf{x}_{\cdot j}\lambda_j$$

but now premultiplying by $\mathbf{x}'_{\cdot i}$

$$\mathbf{x}'_{\cdot i}\mathbf{S}\mathbf{x} = \mathbf{x}'_{\cdot i}\mathbf{x}_{\cdot j}\lambda_j \,.$$

Clearly it has to be the case that

$$\mathbf{x}'_{\cdot j}\mathbf{S}\mathbf{x}_{\cdot i} = \mathbf{x}'_{\cdot i}\mathbf{S}\mathbf{x}_{\cdot j}$$

in which case,

$$\mathbf{x}'_{\cdot j}\mathbf{x}_{\cdot i}\lambda_i = \mathbf{x}'_{\cdot i}\mathbf{x}_{\cdot j}\lambda_j \,.$$

But for that to happen, it must be true that

$$\mathbf{x}'_{\cdot j}\mathbf{x}_{\cdot i} = 0. \tag{3.32}$$

In other words, each pair of eigenvectors is orthogonal. When you add the standardizing constraint, Equation (3.23), we can say that

$$\mathbf{X}'\mathbf{X} = \mathbf{I} . \tag{3.33}$$

The **X** matrix, as can be seen above, acts as its own inverse. Any matrix **X** for which $\mathbf{X}'\mathbf{X} = \mathbf{X}\mathbf{X}' = \mathbf{I}$ is called *orthonormal*.

Here are some more properties of the eigenvalues and eigenvectors. From Equation (3.27) we can make the simultaneous statement about each eigenvalue-eigenvector below,

$$\mathbf{S}\mathbf{X} = \mathbf{X}\mathbf{L}. \tag{3.34}$$

Premultiplying by $\mathbf{X}'$ leads to

$$\mathbf{X}'\mathbf{S}\mathbf{X} = \mathbf{L} . \tag{3.35}$$

Or, starting again from Equation (3.34) but postmultiplying by $\mathbf{X}'$ this time leads to

$$\mathbf{S} = \mathbf{X}\mathbf{L}\mathbf{X}' . \tag{3.36}$$

$$= \mathbf{X}\mathbf{L}^{1/2}\,\mathbf{L}^{1/2}\mathbf{X}' \tag{3.37}$$

where the "square root" of the matrix **L** is clearly defined as $\{\lambda_i^{1/2}\}$, that is having the square root of each of the λ_i on the diagonal [c.f. Equation (2.14) and the discussion thereof]. Now if we define

$$\mathbf{B} = \mathbf{X}\mathbf{L}^{1/2}$$

We can say that

$$\mathbf{S} = \mathbf{B}\mathbf{B}' \tag{3.38}$$

which provides a "square root" like effect, even if the square root of a non-diagonal matrix cannot be uniquely defined. That this equation is not unique can be shown simply by defining the orthonormal matrix **J**, i. e. $\mathbf{J}'\mathbf{J} = \mathbf{J}\mathbf{J}' = \mathbf{I}$. Now if $\mathbf{B}^* = \mathbf{B}\mathbf{J}$ then

$$\mathbf{S} = \mathbf{B}^{*\prime}\,\mathbf{B}^* = \mathbf{B}\mathbf{J}\mathbf{J}'\mathbf{B} = \mathbf{B}\mathbf{B}'.$$

In factor analysis we seek a **B** matrix corresponding to a hypothesis about latent variables. In Cholesky factorization, we produce a lower triangular **B** matrix. In finding the eigenstructure of the **S** matrix, the columns of the **B** matrix produced in Equation 3.38) maximize the variance of the extracted components.

But the eigenstructure of **S** captures even more of the properties of **S**. For example, if $\mathbf{S}^{-1}$ exists,

$$\mathbf{S}^{-1} = \mathbf{X}\mathbf{L}^{-1}\mathbf{X}' . \tag{3.39}$$

In addition, if $\mathbf{A} = c\mathbf{S}$ where c is a scalar, then

$$\mathbf{A} = \mathbf{X}c\mathbf{L}\mathbf{X}' , \tag{3.40}$$

and if $\mathbf{A} = \mathbf{S} + c\mathbf{I}$ then

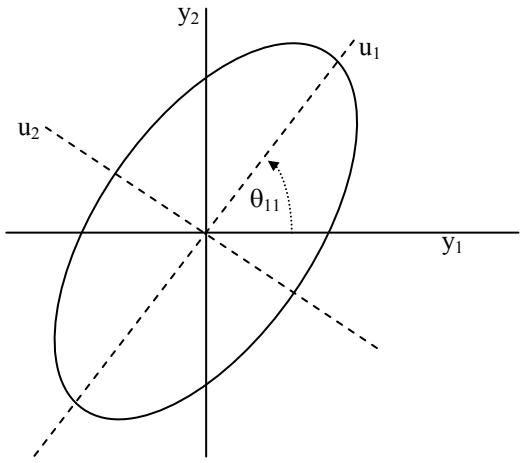
$$\mathbf{A} = \mathbf{X}[\mathbf{L} + c\mathbf{I}]\mathbf{X}' . \tag{3.41}$$

3.8 *Some Geometric Aspects of Eigenstructure*

Since $\mathbf{X}'\mathbf{X} = \mathbf{I}$, $\mathbf{X}$ can be thought of as a rigid, or *angle-preserving transformation* of a coordinate space. The original vector $\mathbf{y}$ is transformed to $\mathbf{u}$ by $\mathbf{X}$ as in

$$\mathbf{u} = \mathbf{X}'y.$$

Here we have repeated Equation (3.21), except the transformation occurs for each eigenvector, not just the first one. Alternatively, instead of thinking of $\mathbf{y}$ as moving to $\mathbf{u}$, we can think of this as the axes of the space moving. A picture of this is now shown:



The angle between an old axis, y_i, and a new axis, u_j, is notated θ_{ij}. We note then for the two dimensional example given above, we have for X

$$\mathbf{X} = \begin{bmatrix} \cos\theta_{11} & \cos\theta_{12} \\ -\cos\theta_{21} & \cos\theta_{22} \end{bmatrix}.$$

The angles θ_{ij} are determined by the direction of the principle axis of the ellipsoid $\mathbf{x}'\mathbf{S}\mathbf{x} = \boldsymbol{\lambda}$.

3.9 *Linear and Nonlinear Parameter Estimation*

In almost all cases that we have in mathematical reasoning in marketing, there are some aspects of our model that we know, for example there might be the value π, and there are some values that we do not know and that therefore have to be estimated from the sample at hand. For example, in the linear model, $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$, the $\mathbf{X}$ matrix is known, but the $\boldsymbol{\beta}$ vector contains a set of regression slopes that need to be estimated from the sample. The topic of linear estimation is investigated in depth in Chapter 5. For now, we note that we create an objective function, that when optimized, will lead us to good estimates for this unknown parameter vector. For example, we might pick the sum of squares of deviations between predicted data and actual data. In that case we would have

$$f = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

$$= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

as our objective function. The goal then is to pick values in the $\boldsymbol{\beta}$ vector so as to make f as small as possible. According to the calculus, this can be done by determining the derivative of f with respect to $\boldsymbol{\beta}$, and setting it equal to zero as in

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

The derivative $\partial f/\partial \boldsymbol{\beta}$ is a linear equation and happens to contain solely elements of the $\mathbf{X}$ matrix, the $\mathbf{y}$ vector and $\boldsymbol{\beta}$ in various combinations. When we set it equal to zero, we can solve for $\boldsymbol{\beta}$ and end up with things on the right hand side that are known, namely $\mathbf{X}$ and $\mathbf{y}$. This allows us to derive a *closed form* or *analytical solution* for $\boldsymbol{\beta}$ that we call $\hat{\boldsymbol{\beta}}$,

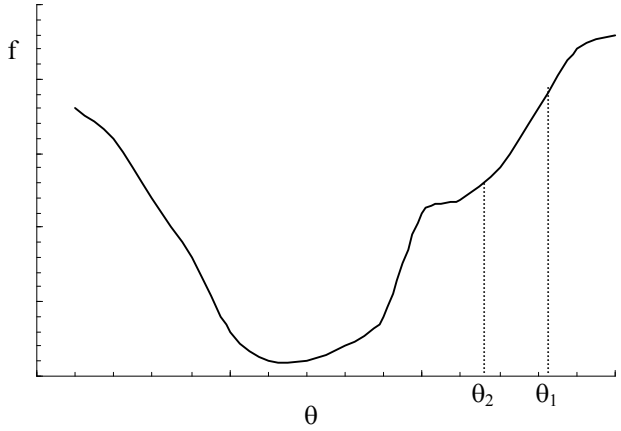
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The term closed-form means that we can use algebraic analysis to find the values for the unknowns. In short, we end up being able to solve for the unknowns. In other cases, our objective function, or its derivative, might be more complex. In those cases we cannot just solve for the unknown parameters using algebra. This often happens when we are trying to model choice probabilities, or market shares, which; since they are bounded by 0 and 1; logically cannot be represented linearly. When this happens we have to use *non-linear optimization*. Non-linear optimization involves the following steps.

1. We take a stab at the unknowns, inventing *starting values* for them and loading them into a vector. Lets call that vector $\boldsymbol{\theta}$.

2. We assess the derivative of the objective function at the current values in $\boldsymbol{\theta}$. If the derivative is not zero, we modify $\boldsymbol{\theta}$ by moving it in the direction in which the derivative is getting closer to $\mathbf{0}$, the null vector. We keep repeating this step until the derivative arrives at the null vector.

How do we know in which direction to move $\boldsymbol{\theta}$? First we will look at a geometric picture, then we will use symbols to make the argument. Lets assume that instead of an entire vector of unknowns, we have a single unknown; the scalar θ. We have started out with an estimate of θ at θ_1.



We are trying to move our estimate towards the bottom of the function. This is logically analogous to a parachutist who lands on the side of the hills surrounding a valley and who wants to find the bottom of the valley in the dead of night. How does he or she know which way to move? By feeling with your foot, you can figure out which way is down. The derivate $\partial f/\partial\theta_1$ gives us the slope of the function that relates θ to f, evaluated at θ_1. It lets us know which way is down. If the derivative is negative, we need to move to our right on the graph, because that is the direction in which f is less. On the other hand, if the derivative is positive, as it would be at position θ_1, we need to move to our left. In more formal terms, in nonlinear optimization we could calculate the next estimate of θ using the formula

$$\theta_{i+1} = \theta_i - \delta\frac{\partial f}{\partial\theta_i}$$

where δ is the step size. Sometimes we use the derivative of the derivative (the second order derivative) to fine-tune the step size. The step size can be important because we want to make sure we end up at the *global minimum* of f, not a *local minimum*. It also can help when you have good, rational, starting values for the first step that are close to their true values. Good start values and a good choice for step size can also make the search go faster, something that is still important even in these days of cheap computing power. In any case, non-linear optimization algorithms stop when the derivative gets close enough to zero, or in other words, when the difference between successive estimates of the unknowns does not change any more. Its important to understand that typically, there are more than one unknown parameters estimated at the same time. Thus the parameters and their derivatives are in vector form.

Nonlinear estimation is used in many branches of statistics and is needed in almost every chapter except for 5, 6, 7 and 8.

3.10 *Maximum Likelihood Parameter Estimation*

Rather than minimize the sum of squared errors, a different philosophy would have us maximize the likelihood of the sample. In general, the probability that our model is correct is proportional to the probability of the data given the model. In *Maximum Likelihood* (ML), we pick parameter estimates such that the probability of the data is as high as possible. Of course, it only makes sense that we would want to maximize the probability of observing the data that we actually did observe.

We can illustrate this using μ, the population mean. Suppose that we had a sample of three people, with scores of 4, 6 and 8. What would the probability be of observing this sample if the true population value of μ was 249? Pretty low, right? What would the probability of the sample be if μ was equal to 6? Certainly it would be quite a bit higher. The likelihood principle tells us to pick that estimate for μ that maximizes the probability of the sample. Of course to do this, we need to make an assumption about the probability distribution of the observations that comprise the sample.

To make the discussion more general, consider a set of observations $y_1, y_2, \cdots, y_n$. Lets say further that we have a model and that the unknown parameters of the model are in the vector $\boldsymbol{\theta}$. According to the model, the likelihood of observation i is $\Pr(y_i \mid \boldsymbol{\theta})$. Assuming independent sample units, i. e. no data point is influenced by any other, the likelihood function according to the model is

$$\ell_0 = \prod_i^n \Pr(y_i \mid \boldsymbol{\theta}) . \tag{3.42}$$

In these cases we also tend to have a version of the $\Pr(y_i)$ that does not depend on $\boldsymbol{\theta}$. The likelihood of the sample under this alternative may be called ℓ_A. It turns out that under very general conditions, $-2\ln(\ell_0/\ell_A)$ is distributed according to the Chi Square distribution, i. e.

$$\hat{\chi}^2 = -2\ln(\ell_0/\ell_A). \tag{3.43}$$

The minus sign in front of the expression for Chi Square means that we can switch from maximizing ℓ_0 to minimizing Chi Square. Minimization is always a safer bet where computers are concerned since a number too large to be processed causes far more of a problem than a number that is too close to zero (the square in Chi Square implies that it is non-negative). What's more, this allows us to test our model against the general alternative hypothesis using the χ^2 distribution. The degrees of freedom of the Chi Square are equal to the difference between the number of data points that we are using; in this case n, and the number of unknown elements in $\boldsymbol{\theta}$.

Here, it could be added that in some cases, such as linear regression, maximum likelihood estimates have a closed form and can be estimated using the formula for $\hat{\boldsymbol{\beta}}$ given in the previous section. In other words, $\hat{\boldsymbol{\beta}}$ does not just minimize the sum of squared errors, it also maximizes the likelihood function. In other cases, we don't get that sort of break and nonlinear optimization must be used.

Maximum likelihood comes with variances and covariances of the parameter vector "built-in". The matrix of the second order derivatives, known as the *Hessian*, contains the elements:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial \hat{\chi}^2}{\partial^2 \theta_1} & \frac{\partial \hat{\chi}^2}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial \hat{\chi}^2}{\partial \theta_1 \partial \theta_q} \\ \frac{\partial \hat{\chi}^2}{\partial \theta_2 \partial \theta_1} & \frac{\partial \hat{\chi}^2}{\partial^2 \theta_2} & \cdots & \frac{\partial \hat{\chi}^2}{\partial \theta_2 \partial \theta_q} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \hat{\chi}^2}{\partial \theta_q \partial \theta_1} & \frac{\partial \hat{\chi}^2}{\partial \theta_q \partial \theta_2} & \cdots & \frac{\partial \hat{\chi}^2}{\partial^2 \theta_q} \end{bmatrix}. \tag{3.44}$$

Elements of the above matrix, $h_{ij} = \frac{\partial \hat{\chi}^2}{\partial \theta_i \partial \theta_j}$, consist of the derivative of the derivative of $\hat{\chi}^2$ with respect to θ_i, with respect to θ_j. In other words,

$$h_{ij} = \frac{\partial}{\partial \theta_i}\left[\frac{\partial \hat{\chi}^2}{\partial \theta_j}\right]. \tag{3.45}$$

Here we are treating $\frac{\partial \hat{\chi}^2}{\partial \theta_j}$ as a function of θ_j, and taking <u>its</u> derivative with respect to θ_j.

The covariance matrix of $\boldsymbol{\theta}$ is given by

$$V(\boldsymbol{\theta}) = [-E(\mathbf{H})]^{-1} \tag{3.46}$$

with the term in the square brackets, -E(**H**), minus the expectation of the Hessian, called the *information matrix*.

Whenever possible, marketing scientists prefer to work with maximum likelihood estimators given that they have very desirable properties. In addition to knowing the variance matrix of your estimator, if $\hat{\theta}$ is the maximum likelihood estimator of θ then $f(\hat{\theta})$ estimates $f(\theta)$ (for more detail see Johnson and Wichern, 2002, p. 170). You can estimate $\hat{\theta}$ and then apply the function f. More importantly, if you can derive or create a maximum likelihood estimator in a certain situation, that estimator is guaranteed to be consistent, asymptotically normally distributed and asymptotically efficient (a proof of this appears in Theil 1971, pp. 392-7). The phrase asymptotically efficient implies that no other estimator can have a lower variance.

References

Johnson, Richard A. and Dean W. Wichern (2002) *Applied Multivariate Statistical Analysis, Fifth Edition*. Upper Saddle River, NJ: Prentice-Hall.

Theil, Henri (1971) *Principles of Econometrics*. New York: John Wiley.