4

Vector/Matrix Derivatives and Integrals

The operations of differentiation and integration of vectors and matrices are logical extensions of the corresponding operations on scalars. There are three objects involved in this operation:

- the variable of the operation;
- the operand (the function being differentiated or integrated); and
- the result of the operation.

In the simplest case, all three of these objects are of the same type, and they are scalars. If either the variable or the operand is a vector or a matrix, however, the structure of the result may be more complicated. This statement will become clearer as we proceed to consider specific cases.

In this chapter, we state or show the form that the derivative takes in terms of simpler derivatives. We state high-level rules for the nature of the differentiation in terms of simple partial differentiation of a scalar with respect to a scalar. We do not consider whether or not the derivatives exist. In general, if the simpler derivatives we write that comprise the more complicated object exist, then the derivative of that more complicated object exists. Once a shape of the derivative is determined, definitions or derivations in ϵ-δ terms could be given, but we will refrain from that kind of formal exercise. The purpose of this chapter is not to develop a calculus for vectors and matrices but rather to consider some cases that find wide applications in statistics. For a more careful treatment of differentiation of vectors and matrices, the reader is referred to Rogers (1980) or to Magnus and Neudecker (1999). Anderson (2003), Muirhead (1982), and Nachbin (1965) cover various aspects of integration with respect to vector or matrix differentials.

4.1 Basics of Differentiation

It is useful to recall the heuristic interpretation of a derivative. A derivative of a function is the infinitesimal rate of change of the function with respect to the variable with which the differentiation is taken. If both the function and the variable are scalars, this interpretation is unambiguous. If, however, the operand of the differentiation, $\mathit{\Phi}$, is a more complicated function, say a vector or a matrix, and/or the variable of the differentiation, $\mathit{\Xi}$, is a more complicated object, the changes are more difficult to measure. Change in the value both of the function,

$$\delta\mathit{\Phi} = \mathit{\Phi}_{\mathrm{new}} - \mathit{\Phi}_{\mathrm{old}},$$

and of the variable,

$$\delta\mathit{\Xi} = \mathit{\Xi}_{\mathrm{new}} - \mathit{\Xi}_{\mathrm{old}},$$

could be measured in various ways; for example, by using various norms, as discussed in Sections 2.1.5 and 3.9. (Note that the subtraction is not necessarily ordinary scalar subtraction.)

Furthermore, we cannot just divide the function values by $\delta\mathit{\Xi}$. We do not have a definition for division by that kind of object. We need a mapping, possibly a norm, that assigns a positive real number to $\delta\mathit{\Xi}$. We can define the change in the function value as just the simple difference of the function evaluated at the two points. This yields

$$\lim_{\|\delta\mathit{\Xi}\|\to 0} \frac{\mathit{\Phi}(\mathit{\Xi}+\delta\mathit{\Xi}) - \mathit{\Phi}(\mathit{\Xi})}{\|\delta\mathit{\Xi}\|}. \tag{4.1}$$

So long as we remember the complexity of $\delta\mathit{\Xi}$, however, we can adopt a simpler approach. Since for both vectors and matrices, we have definitions of multiplication by a scalar and of addition, we can simplify the limit in the usual definition of a derivative, $\delta\mathit{\Xi} \to 0$. Instead of using $\delta\mathit{\Xi}$ as the element of change, we will use $t\mathit{\Upsilon}$, where t is a scalar and $\mathit{\Upsilon}$ is an element to be added to $\mathit{\Xi}$. The limit then will be taken in terms of $t \to 0$. This leads to

$$\lim_{t\to 0} \frac{\mathit{\Phi}(\mathit{\Xi}+t\mathit{\Upsilon}) - \mathit{\Phi}(\mathit{\Xi})}{t} \tag{4.2}$$

as a formula for the derivative of $\mathit{\Phi}$ with respect to $\mathit{\Xi}$.

The expression (4.2) may be a useful formula for evaluating a derivative, but we must remember that it is not the derivative. The type of object of this formula is the same as the type of object of the function, $\mathit{\Phi}$; it does not accommodate the type of object of the argument, $\mathit{\Xi}$, unless $\mathit{\Xi}$ is a scalar. As we will see below, for example, if $\mathit{\Xi}$ is a vector and $\mathit{\Phi}$ is a scalar, the derivative must be a vector, yet in that case the expression (4.2) is a scalar.

The expression (4.1) is rarely directly useful in evaluating a derivative, but it serves to remind us of both the generality and the complexity of the concept. Both $\mathit{\Phi}$ and its arguments could be functions, for example. (In functional analysis, various kinds of functional derivatives are defined, such as a Gâteaux derivative. These derivatives find applications in developing robust statistical methods; see Shao, 2003, for example.) In this chapter, we are interested in the combinations of three possibilities for $\mathit{\Phi}$, namely scalar, vector, and matrix, and the same three possibilities for $\mathit{\Xi}$ and $\mathit{\Upsilon}$.

Continuity

It is clear from the definition of continuity that for the derivative of a function to exist at a point, the function must be continuous at that point. A function of a vector or a matrix is continuous if it is continuous for each element of the vector or matrix. Just as scalar sums and products are continuous, vector/matrix sums and all of the types of vector/matrix products we have discussed are continuous. A continuous function of a continuous function is continuous.

Many of the vector/matrix functions we have discussed are clearly continuous. For example, the L_p vector norms in equation (2.11) are continuous over the nonnegative reals but not over the reals unless p is an even (positive) integer. The determinant of a matrix is continuous, as we see from the definition of the determinant and the fact that sums and scalar products are continuous. The fact that the determinant is a continuous function immediately yields the result that cofactors and hence the adjugate are continuous. From the relationship between an inverse and the adjugate (equation (3.131)), we see that the inverse is a continuous function.

Notation and Properties

We write the differential operator with respect to the dummy variable x as $\partial/\partial x$ or $\partial/\partial x^{\mathrm{T}}$. We usually denote differentiation using the symbol for "partial" differentiation, ∂, whether the operator is written ∂x_i for differentiation with respect to a specific scalar variable or ∂x for differentiation with respect to the array x that contains all of the individual elements. Sometimes, however, if the differentiation is being taken with respect to the whole array (the vector or the matrix), we use the notation $\mathrm{d}/\mathrm{d}x$.

The operand of the differential operator $\partial/\partial x$ is a function of x. (If it is not a function of x — that is, if it is a constant function with respect to x — then the operator evaluates to 0.) The result of the operation, written $\partial f/\partial x$, is also a function of x, with the same domain as f, and we sometimes write $\partial f(x)/\partial x$ to emphasize this fact. The value of this function at the fixed point x_0 is written as $\partial f(x_0)/\partial x$. (The derivative of the constant $f(x_0)$ is identically 0, but it is not necessary to write $\partial f(x)/\partial x|_{x_0}$ because $\partial f(x_0)/\partial x$ is interpreted as the value of the function $\partial f(x)/\partial x$ at the fixed point x_0.)

If $\partial/\partial x$ operates on f, and $f : S \to T$, then $\partial/\partial x : S \to U$. The nature of S, or more directly the nature of x, whether it is a scalar, a vector, or a matrix, and the nature of T determine the structure of the result U. For example, if x is an n-vector and $f(x) = x^{\mathrm{T}}x$, then

$$f : \mathbb{R}^n \to \mathbb{R}$$

and

$$\partial f/\partial x : \mathbb{R}^n \to \mathbb{R}^n,$$

as we will see. The outer product, $h(x) = xx^{\mathrm{T}}$, is a mapping to a higher rank array, but the derivative of the outer product is a mapping to an array of the same rank; that is,

$$h : \mathrm{I\!R}^n \to \mathrm{I\!R}^{n\times n}$$

and

$$\partial h/\partial x : \mathrm{I\!R}^n \to \mathrm{I\!R}^n.$$

(Note that "rank" here means the number of dimensions; see page 5.)

As another example, consider $g(\cdot) = \det(\cdot)$, so

$$g : \mathrm{I\!R}^{n\times n} \mapsto \mathrm{I\!R}.$$

In this case,

$$\partial g/\partial X : \mathrm{I\!R}^{n\times n} \mapsto \mathrm{I\!R}^{n\times n};$$

that is, the derivative of the determinant of a square matrix is a square matrix, as we will see later.

Higher-order differentiation is a composition of the $\partial/\partial x$ operator with itself or of the $\partial/\partial x$ operator and the $\partial/\partial x^{\mathrm{T}}$ operator. For example, consider the familiar function in linear least squares

$$f(b) = (y - Xb)^{\mathrm{T}}(y - Xb).$$

This is a mapping from $\mathrm{I\!R}^m$ to $\mathrm{I\!R}$. The first derivative with respect to the m-vector b is a mapping from $\mathrm{I\!R}^m$ to $\mathrm{I\!R}^m$, namely $2X^{\mathrm{T}}Xb - 2X^{\mathrm{T}}y$. The second derivative with respect to b^{T} is a mapping from $\mathrm{I\!R}^m$ to $\mathrm{I\!R}^{m\times m}$, namely, $2X^{\mathrm{T}}X$. (Many readers will already be familiar with these facts. We will discuss the general case of differentiation with respect to a vector in Section 4.2.2.)

We see from expression (4.1) that differentiation is a linear operator; that is, if $\mathcal{D}(\Phi)$ represents the operation defined in expression (4.1), Ψ is another function in the class of functions over which $\mathcal{D}$ is defined, and a is a scalar that does not depend on the variable Ξ, then $\mathcal{D}(a\Phi + \Psi) = a\mathcal{D}(\Phi) + \mathcal{D}(\Psi)$. This yields the familiar rules of differential calculus for derivatives of sums or constant scalar products. Other usual rules of differential calculus apply, such as for differentiation of products and composition (the chain rule). We can use expression (4.2) to work these out. For example, for the derivative of the product $\Phi\Psi$, after some rewriting of terms, we have the numerator

$$\begin{aligned}
&\Phi(\Xi)\big(\Psi(\Xi + t\Upsilon) - \Psi(\Xi)\big)\\
&\quad+\Psi(\Xi)\big(\Phi(\Xi + t\Upsilon) - \Phi(\Xi)\big)\\
&\quad+\big(\Phi(\Xi + t\Upsilon) - \Phi(\Xi)\big)\big(\Psi(\Xi + t\Upsilon) - \Psi(\Xi)\big).
\end{aligned}$$

Now, dividing by t and taking the limit, assuming that as

$$t \to 0,$$

$$(\Phi(\Xi + t\Upsilon) - \Phi(\Xi)) \to 0,$$

we have

$$\mathcal{D}(\Phi\Psi) = \mathcal{D}(\Phi)\Psi + \Phi\mathcal{D}(\Psi), \tag{4.3}$$

where again $\mathcal{D}$ represents the differentiation operation.

Differentials

For a differentiable scalar function of a scalar variable, $f(x)$, the *differential of f at c with increment u* is $u\mathrm{d}f/\mathrm{d}x|_c$. This is the linear term in a truncated Taylor series expansion:

$$f(c+u) = f(c) + u\frac{\mathrm{d}}{\mathrm{d}x}f(c) + r(c,u). \tag{4.4}$$

Technically, the differential is a function of both x and u, but the notation $\mathrm{d}f$ is used in a generic sense to mean the differential of f. For vector/matrix functions of vector/matrix variables, the differential is defined in a similar way. The structure of the differential is the same as that of the function; that is, for example, the differential of a matrix-valued function is a matrix.

4.2 Types of Differentiation

In the following sections we consider differentiation with respect to different types of objects first, and we consider differentiation of different types of objects.

4.2.1 Differentiation with Respect to a Scalar

Differentiation of a structure (vector or matrix, for example) with respect to a scalar is quite simple; it just yields the ordinary derivative of each element of the structure in the same structure. Thus, the derivative of a vector or a matrix with respect to a scalar variable is a vector or a matrix, respectively, of the derivatives of the individual elements.

Differentiation with respect to a vector or matrix, which we will consider below, is often best approached by considering differentiation with respect to the individual elements of the vector or matrix, that is, with respect to scalars.

Derivatives of Vectors with Respect to Scalars

The derivative of the vector $y(x) = (y_1, \ldots, y_n)$ with respect to the scalar x is the vector

$$\partial y/\partial x = (\partial y_1/\partial x, \ldots, \partial y_n/\partial x). \tag{4.5}$$

The second or higher derivative of a vector with respect to a scalar is likewise a vector of the derivatives of the individual elements; that is, it is an array of higher rank.

Derivatives of Matrices with Respect to Scalars

The derivative of the matrix $Y(x) = (y_{ij})$ with respect to the scalar x is the matrix

$$\partial Y(x)/\partial x = (\partial y_{ij}/\partial x). \tag{4.6}$$

The second or higher derivative of a matrix with respect to a scalar is likewise a matrix of the derivatives of the individual elements.

Derivatives of Functions with Respect to Scalars

Differentiation of a function of a vector or matrix that is linear in the elements of the vector or matrix involves just the differentiation of the elements, followed by application of the function. For example, the derivative of a trace of a matrix is just the trace of the derivative of the matrix. On the other hand, the derivative of the determinant of a matrix is not the determinant of the derivative of the matrix (see below).

Higher-Order Derivatives with Respect to Scalars

Because differentiation with respect to a scalar does not change the rank of the object ("rank" here means rank of an array or "shape"), higher-order derivatives $\partial^k/\partial x^k$ with respect to scalars are merely objects of the same rank whose elements are the higher-order derivatives of the individual elements.

4.2.2 Differentiation with Respect to a Vector

Differentiation of a given object with respect to an n-vector yields a vector for each element of the given object. The basic expression for the derivative, from formula (4.2), is

$$\lim_{t\to 0} \frac{\Phi(x+ty) - \Phi(x)}{t} \tag{4.7}$$

for an arbitrary conformable vector y. The arbitrary y indicates that the derivative is omnidirectional; it is the rate of change of a function of the vector in any direction.

Derivatives of Scalars with Respect to Vectors; The Gradient

The derivative of a scalar-valued function with respect to a vector is a vector of the partial derivatives of the function with respect to the elements of the vector. If $f(x)$ is a scalar function of the vector $x = (x_1, \ldots, x_n)$,

$$\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right), \tag{4.8}$$

if those derivatives exist. This vector is called the *gradient* of the scalar-valued function, and is sometimes denoted by $\mathrm{g}_f(x)$ or $\nabla f(x)$, or sometimes just g_f or ∇f:

$$\mathrm{g}_f = \nabla f = \frac{\partial f}{\partial x}. \tag{4.9}$$

The notation g_f or ∇f implies differentiation with respect to "all" arguments of f, hence, if f is a scalar-valued function of a vector argument, they represent a vector.

This derivative is useful in finding the maximum or minimum of a function. Such applications arise throughout statistical and numerical analysis. In Section 6.3.2, we will discuss a method of solving linear systems of equations by formulating the problem as a minimization problem.

Inner products, bilinear forms, norms, and variances are interesting scalar-valued functions of vectors. In these cases, the function Φ in equation (4.7) is scalar-valued and the numerator is merely $\Phi(x + ty) - \Phi(x)$. Consider, for example, the quadratic form $x^{\mathrm{T}}Ax$. Using equation (4.7) to evaluate $\partial x^{\mathrm{T}}Ax/\partial x$, we have

$$\begin{aligned}
&\lim_{t\to 0} \frac{(x + ty)^{\mathrm{T}}A(x + ty) - x^{\mathrm{T}}Ax}{t} \\
&= \lim_{t\to 0} \frac{x^{\mathrm{T}}Ax + ty^{\mathrm{T}}Ax + ty^{\mathrm{T}}A^{\mathrm{T}}x + t^2y^{\mathrm{T}}Ay - x^{\mathrm{T}}Ax}{t} \\
&= y^{\mathrm{T}}(A + A^{\mathrm{T}})x,
\end{aligned} \tag{4.10}$$

for an arbitrary y (that is, "in any direction"), and so $\partial x^{\mathrm{T}}Ax/\partial x = (A+A^{\mathrm{T}})x$.

This immediately yields the derivative of the square of the Euclidean norm of a vector, $\|x\|_2^2$, and the derivative of the Euclidean norm itself by using the chain rule. Other L_p vector norms may not be differentiable everywhere because of the presence of the absolute value in their definitions. The fact that the Euclidean norm is differentiable everywhere is one of its most important properties.

The derivative of the quadratic form also immediately yields the derivative of the variance. The derivative of the correlation, however, is slightly more difficult because it is a ratio (see Exercise 4.2).

The operator $\partial/\partial x^{\mathrm{T}}$ applied to the scalar function f results in $\mathrm{g}_f^{\mathrm{T}}$.

The second derivative of a scalar-valued function with respect to a vector is a derivative of the first derivative, which is a vector. We will now consider derivatives of vectors with respect to vectors.

Derivatives of Vectors with Respect to Vectors; The Jacobian

The derivative of an m-vector-valued function of an n-vector argument consists of nm scalar derivatives. These derivatives could be put into various structures. Two obvious structures are an $n \times m$ matrix and an $m \times n$ matrix. For a function $f : S \subset \mathrm{I\!R}^n \to \mathrm{I\!R}^m$, we define $\partial f^{\mathrm{T}}/\partial x$ to be the $n \times m$ matrix, which is the natural extension of $\partial/\partial x$ applied to a scalar function, and $\partial f/\partial x^{\mathrm{T}}$ to be its transpose, the $m \times n$ matrix. Although the notation $\partial f^{\mathrm{T}}/\partial x$ is more precise because it indicates that the elements of f correspond to the columns of the result, we often drop the transpose in the notation. We have

$$\begin{aligned}
\frac{\partial f}{\partial x} &= \frac{\partial f^{\mathrm{T}}}{\partial x} \quad \text{by convention} \\
&= \left[\frac{\partial f_1}{\partial x} \cdots \frac{\partial f_m}{\partial x} \right] \\
&= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ & & \cdots & \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}
\end{aligned} \tag{4.11}$$

if those derivatives exist. This derivative is called the *matrix gradient* and is denoted by G_f or ∇f for the vector-valued function f. (Note that the ∇ symbol can denote either a vector or a matrix, depending on whether the function being differentiated is scalar-valued or vector-valued.)

The $m \times n$ matrix $\partial f/\partial x^{\mathrm{T}} = (\nabla f)^{\mathrm{T}}$ is called the *Jacobian* of f and is denoted by J_f:

$$\mathrm{J}_f = \mathrm{G}_f^{\mathrm{T}} = (\nabla f)^{\mathrm{T}}. \tag{4.12}$$

The absolute value of the determinant of the Jacobian appears in integrals involving a change of variables. (Occasionally, the term "Jacobian" is used to refer to the absolute value of the determinant rather than to the matrix itself.)

To emphasize that the quantities are functions of x, we sometimes write $\partial f(x)/\partial x$, $\mathrm{J}_f(x)$, $\mathrm{G}_f(x)$, or $\nabla f(x)$.

Derivatives of Matrices with Respect to Vectors

The derivative of a matrix with respect to a vector is a three-dimensional object that results from applying equation (4.8) to each of the elements of the matrix. For this reason, it is simpler to consider only the partial derivatives of the matrix Y with respect to the individual elements of the vector x; that is, $\partial Y/\partial x_i$. The expressions involving the partial derivatives can be thought of as defining one two-dimensional layer of a three-dimensional object.

Using the rules for differentiation of powers that result directly from the definitions, we can write the partial derivatives of the inverse of the matrix Y as

$$\frac{\partial}{\partial x} Y^{-1} = -Y^{-1} \left(\frac{\partial}{\partial x} Y \right) Y^{-1} \tag{4.13}$$

(see Exercise 4.3).

Beyond the basics of differentiation of constant multiples or powers of a variable, the two most important properties of derivatives of expressions are the linearity of the operation and the chaining of the operation. These yield rules that correspond to the familiar rules of the differential calculus. A simple result of the linearity of the operation is the rule for differentiation of the trace:

$$\frac{\partial}{\partial x}\mathrm{tr}(Y) = \mathrm{tr}\left(\frac{\partial}{\partial x}Y\right).$$

Higher-Order Derivatives with Respect to Vectors; The Hessian

Higher-order derivatives are derivatives of lower-order derivatives. As we have seen, a derivative of a given function with respect to a vector is a more complicated object than the original function. The simplest higher-order derivative with respect to a vector is the second-order derivative of a scalar-valued function. Higher-order derivatives may become uselessly complicated.

In accordance with the meaning of derivatives of vectors with respect to vectors, the second derivative of a scalar-valued function with respect to a vector is a matrix of the partial derivatives of the function with respect to the elements of the vector. This matrix is called the *Hessian*, and is denoted by H_f or sometimes by $\nabla\nabla f$ or $\nabla^2 f$:

$$\mathrm{H}_f = \frac{\partial^2 f}{\partial x \partial x^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_m} \\ & \cdots & & \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \frac{\partial^2 f}{\partial x_m \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix}. \tag{4.14}$$

To emphasize that the Hessian is a function of x, we sometimes write $\mathrm{H}_f(x)$ or $\nabla\nabla f(x)$ or $\nabla^2 f(x)$.

Summary of Derivatives with Respect to Vectors

As we have seen, the derivatives of functions are complicated by the problem of measuring the change in the function, but often the derivatives of functions with respect to a vector can be determined by using familiar scalar differentiation. In general, we see that

- the derivative of a scalar (a quadratic form) with respect to a vector is a vector and
- the derivative of a vector with respect to a vector is a matrix.

Table 4.1 lists formulas for the vector derivatives of some common expressions. The derivative $\partial f/\partial x^{\mathrm{T}}$ is the transpose of $\partial f/\partial x$.

Table 4.1. Formulas for Some Vector Derivatives

$f(x)$	$\partial f/\partial x$
ax	aI
$b^{\mathrm{T}}x$	b
$x^{\mathrm{T}}b$	b^{T}
$x^{\mathrm{T}}x$	$I \otimes x + x \otimes I$
xx^{T}	$2x^{\mathrm{T}}$
$b^{\mathrm{T}}Ax$	$A^{\mathrm{T}}b$
$x^{\mathrm{T}}Ab$	$b^{\mathrm{T}}A$
$x^{\mathrm{T}}Ax$	$(A + A^{\mathrm{T}})x$
	$2Ax$, if A is symmetric
$\exp(-\frac{1}{2}x^{\mathrm{T}}Ax)$	$-\exp(-\frac{1}{2}x^{\mathrm{T}}Ax)Ax$, if A is symmetric
$\Vert x\Vert_2^2$	$2x$
$\mathrm{V}(x)$	$2x/(n-1)$

In this table, x is an n-vector, a is a constant scalar, b is a constant conformable vector, and A is a constant conformable matrix.

4.2.3 Differentiation with Respect to a Matrix

The derivative of a function with respect to a matrix is a matrix with the same shape consisting of the partial derivatives of the function with respect to the elements of the matrix. This rule defines what we mean by differentiation with respect to a matrix.

By the definition of differentiation with respect to a matrix X, we see that the derivative $\partial f/\partial X^{\mathrm{T}}$ is the transpose of $\partial f/\partial X$. For scalar-valued functions, this rule is fairly simple. For example, consider the trace. If X is a square matrix and we apply this rule to evaluate $\partial \operatorname{tr}(X)/\partial X$, we get the identity matrix, where the nonzero elements arise only when $j = i$ in $\partial(\sum x_{ii})/\partial x_{ij}$. If AX is a square matrix, we have for the (i, j) term in $\partial \operatorname{tr}(AX)/\partial X$, $\partial \sum_i \sum_k a_{ik}x_{ki}/\partial x_{ij} = a_{ji}$, and so $\partial \operatorname{tr}(AX)/\partial X = A^{\mathrm{T}}$, and likewise, inspecting $\partial \sum_i \sum_k x_{ik}x_{ki}/\partial x_{ij}$, we get $\partial \operatorname{tr}(X^{\mathrm{T}}X)/\partial X = 2X^{\mathrm{T}}$. Likewise for the scalar-valued $a^{\mathrm{T}}Xb$, where a and b are conformable constant vectors, for $\partial \sum_m(\sum_k a_k x_{km})b_m/\partial x_{ij} = a_i b_j$, so $\partial a^{\mathrm{T}}Xb/\partial X = ab^{\mathrm{T}}$.

Now consider $\partial|X|/\partial X$. Using an expansion in cofactors (equation (3.21) or (3.22)), the only term in $|X|$ that involves x_{ij} is $x_{ij}(-1)^{i+j}|X_{-(i)(j)}|$, and the cofactor $(x_{(ij)}) = (-1)^{i+j}|X_{-(i)(j)}|$ does not involve x_{ij}. Hence, $\partial|X|/\partial x_{ij} = (x_{(ij)})$, and so $\partial|X|/\partial X = (\operatorname{adj}(X))^{\mathrm{T}}$ from equation (3.24). Using equation (3.131), we can write this as $\partial|X|/\partial X = |X|X^{-\mathrm{T}}$.

The chain rule can be used to evaluate $\partial \log|X|/\partial X$.

Applying the rule stated at the beginning of this section, we see that the derivative of a matrix Y with respect to the matrix X is

$$\frac{\mathrm{d}Y}{\mathrm{d}X} = \frac{\mathrm{d}\mathrm{vec}(Y))}{\mathrm{d}\mathrm{vec}(X))} \quad \left(\text{or } = Y \otimes \frac{\mathrm{d}}{\mathrm{d}X} \text{ especially in statistical applications}\right). \tag{4.15}$$

Table 4.2 lists some formulas for the matrix derivatives of some common expressions. The derivatives shown in Table 4.2 can be obtained by evaluating expression (4.15), possibly also using the chain rule.

Table 4.2. Formulas for Some Matrix Derivatives

General X

$f(X)$	$\partial f/\partial X$
$a^{\mathrm{T}}Xb$	ab^{T}
$\mathrm{tr}(AX)$	A^{T}
$\mathrm{tr}(X^{\mathrm{T}}X)$	$2X^{\mathrm{T}}$
BX	$I_n \otimes B$
XC	$C^{\mathrm{T}} \otimes I_m$
BXC	$C^{\mathrm{T}} \otimes B$

Square and Possibly Invertible X

$f(X)$	$\partial f/\partial X$
$\mathrm{tr}(X)$	I_n
$\mathrm{tr}(X^k)$	kX^{k-1}
$\mathrm{tr}(BX^{-1}C)$	$-(X^{-1}CBX^{-1})^{\mathrm{T}}$
$\|X\|$	$\|X\|X^{-\mathrm{T}}$
$\log\|X\|$	$X^{-\mathrm{T}}$
$\|X\|^k$	$k\|X\|^kX^{-\mathrm{T}}$
$BX^{-1}C$	$-(X^{-1}C)^{\mathrm{T}} \otimes BX^{-1}$

In this table, X is an $n \times m$ matrix, a is a constant n-vector, b is a constant m-vector, A is a constant $m \times n$ matrix, B is a constant $p \times n$ matrix, and C is a constant $m \times q$ matrix.

There are some interesting applications of differentiation with respect to a matrix in maximum likelihood estimation. Depending on the structure of the parameters in the distribution, derivatives of various types of objects may be required. For example, the determinant of a variance-covariance matrix, in the sense that it is a measure of a volume, often occurs as a normalizing factor in a probability density function; therefore, we often encounter the need to differentiate a determinant with respect to a matrix.

4.3 Optimization of Functions

Because a derivative measures the rate of change of a function, a point at which the derivative is equal to 0 is a stationary point, which may be a maximum or a minimum of the function. Differentiation is therefore a very useful tool for finding the optima of functions, and so, for a given function $f(x)$, the gradient vector function, $\mathrm{g}_f(x)$, and the Hessian matrix function, $\mathrm{H}_f(x)$, play important roles in optimization methods.

We may seek either a maximum or a minimum of a function. Since maximizing the scalar function $f(x)$ is equivalent to minimizing $-f(x)$, we can always consider optimization of a function to be minimization of a function. Thus, we generally use terminology for the problem of finding a minimum of a function. Because the function may have many ups and downs, we often use the phrase *local minimum* (or local maximum or local optimum).

Except in the very simplest of cases, the optimization method must be iterative, moving through a sequence of points, $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$, that approaches the optimum point arbitrarily closely. At the point $x^{(k)}$, the direction of *steepest descent* is clearly $-\mathrm{g}_f(x^{(k)})$, but because this direction may be continuously changing, the steepest descent direction may not be the best direction in which to seek the next point, $x^{(k+1)}$.

4.3.1 Stationary Points of Functions

The first derivative helps only in finding a stationary point. The matrix of second derivatives, the Hessian, provides information about the nature of the stationary point, which may be a local minimum or maximum, a saddlepoint, or only an inflection point.

The so-called second-order optimality conditions are the following (see a general text on optimization for their proofs).

- If (but not only if) the stationary point is a local minimum, then the Hessian is nonnegative definite.
- If the Hessian is positive definite, then the stationary point is a local minimum.
- Likewise, if the stationary point is a local maximum, then the Hessian is nonpositive definite, and if the Hessian is negative definite, then the stationary point is a local maximum.
- If the Hessian has both positive and negative eigenvalues, then the stationary point is a saddlepoint.

4.3.2 Newton's Method

We consider a differentiable scalar-valued function of a vector argument, $f(x)$. By a Taylor series about a stationary point x_*, truncated after the second-order term

$$f(x) \approx f(x_*) + (x - x_*)^{\mathrm{T}} \mathrm{g}_f\big(x_*\big) + \frac{1}{2}(x - x_*)^{\mathrm{T}} \mathrm{H}_f\big(x_*\big)(x - x_*), \tag{4.16}$$

because $\mathrm{g}_f\big(x_*\big) = 0$, we have a general method of finding a stationary point for the function $f(\cdot)$, called Newton's method. If x is an m-vector, $\mathrm{g}_f(x)$ is an m-vector and $\mathrm{H}_f(x)$ is an $m \times m$ matrix.

Newton's method is to choose a starting point $x^{(0)}$, then, for $k = 0, 1, \ldots$, to solve the linear systems

$$\mathrm{H}_f\big(x^{(k)}\big)p^{(k+1)} = -\mathrm{g}_f\big(x^{(k)}\big) \tag{4.17}$$

for $p^{(k+1)}$, and then to update the point in the domain of $f(\cdot)$ by

$$x^{(k+1)} = x^{(k)} + p^{(k+1)}. \tag{4.18}$$

The two steps are repeated until there is essentially no change from one iteration to the next. If $f(\cdot)$ is a quadratic function, the solution is obtained in one iteration because equation (4.16) is exact. These two steps have a very simple form for a function of one variable (see Exercise 4.4a).

Linear Least Squares

In a least squares fit of a linear model

$$y = X\beta + \epsilon, \tag{4.19}$$

where y is an n-vector, X is an $n \times m$ matrix, and β is an m-vector, we replace β by a variable b, define the residual vector

$$r = y - Xb, \tag{4.20}$$

and minimize its Euclidean norm,

$$f(b) = r^{\mathrm{T}} r, \tag{4.21}$$

with respect to the variable b. We can solve this optimization problem by taking the derivative of this sum of squares and equating it to zero. Doing this, we get

$$\begin{aligned}\frac{\mathrm{d}(y - Xb)^{\mathrm{T}}(y - Xb)}{\mathrm{d}b} &= \frac{\mathrm{d}(y^{\mathrm{T}}y - 2b^{\mathrm{T}}X^{\mathrm{T}}y + b^{\mathrm{T}}X^{\mathrm{T}}Xb)}{\mathrm{d}b} \\ &= -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}Xb \\ &= 0,\end{aligned}$$

which yields the normal equations

$$X^{\mathrm{T}}Xb = X^{\mathrm{T}}y.$$

The solution to the normal equations is a stationary point of the function (4.21). The Hessian of $(y - Xb)^{\mathrm{T}}(y - Xb)$ with respect to b is $2X^{\mathrm{T}}X$ and

$$X^{\mathrm{T}}X \succeq 0.$$

Because the matrix of second derivatives is nonnegative definite, the value of b that solves the system of equations arising from the first derivatives is a local minimum of equation (4.21). We discuss these equations further in Sections 6.7 and 9.2.2.

Quasi-Newton Methods

All gradient-descent methods determine the path $p^{(k)}$ to take in the k^{th} step by a system of equations of the form

$$R^{(k)}p^{(k)} = -\mathrm{g}_f\big(x^{(k-1)}\big).$$

In the steepest-descent method, $R^{(k)}$ is the identity, I, in these equations. For functions with eccentric contours, the steepest-descent method traverses a zigzag path to the minimum. In Newton's method, $R^{(k)}$ is the Hessian evaluated at the previous point, $\mathrm{H}_f\big(x^{(k-1)}\big)$, which results in a more direct path to the minimum. Aside from the issues of consistency of the resulting equation and the general problems of reliability, a major disadvantage of Newton's method is the computational burden of computing the Hessian, which requires $\mathrm{O}(m^2)$ function evaluations, and solving the system, which requires $\mathrm{O}(m^3)$ arithmetic operations, at each iteration.

Instead of using the Hessian at each iteration, we may use an approximation, $B^{(k)}$. We may choose approximations that are simpler to update and/or that allow the equations for the step to be solved more easily. Methods using such approximations are called *quasi-Newton* methods or *variable metric* methods.

Because

$$\mathrm{H}_f\big(x^{(k)}\big)\big(x^{(k)} - x^{(k-1)}\big) \approx \mathrm{g}_f\big(x^{(k)}\big) - \mathrm{g}_f\big(x^{(k-1)}\big),$$

we choose $B^{(k)}$ so that

$$B^{(k)}\big(x^{(k)} - x^{(k-1)}\big) = \mathrm{g}_f\big(x^{(k)}\big) - \mathrm{g}_f\big(x^{(k-1)}\big). \qquad (4.22)$$

This is called the *secant condition.*

We express the secant condition as

$$B^{(k)}s^{(k)} = y^{(k)}, \qquad (4.23)$$

where

$$s^{(k)} = x^{(k)} - x^{(k-1)}$$

and

$$y^{(k)} = \mathrm{g}_f(x^{(k)}) - \mathrm{g}_f(x^{(k-1)}),$$

as above.

The system of equations in (4.23) does not fully determine $B^{(k)}$ of course. Because $B^{(k)}$ should approximate the Hessian, we may require that it be symmetric and positive definite.

The most common approach in quasi-Newton methods is first to choose a reasonable starting matrix $B^{(0)}$ and then to choose subsequent matrices by additive updates,

$$B^{(k+1)} = B^{(k)} + B_a^{(k)}, \tag{4.24}$$

subject to preservation of symmetry and positive definiteness. An approximate Hessian $B^{(k)}$ may be used for several iterations before it is updated; that is, $B_a^{(k)}$ may be taken as 0 for several successive iterations.

4.3.3 Optimization of Functions with Restrictions

Instead of the simple least squares problem of determining a value of b that minimizes the sum of squares, we may have some restrictions that b must satisfy; for example, we may have the requirement that the elements of b sum to 1. More generally, consider the least squares problem for the linear model (4.19) with the requirement that b satisfy some set of linear restrictions, $Ab = c$, where A is a full-rank $k \times m$ matrix (with $k \leq m$). (The rank of A must be less than m or else the constraints completely determine the solution to the problem. If the rank of A is less than k, however, some rows of A and some elements of b could be combined into a smaller number of constraints. We can therefore assume A is of full row rank. Furthermore, we assume the linear system is consistent (that is, $\mathrm{rank}([A|c]) = k$) for otherwise there could be no solution.) We call any point b that satisfies $Ab = c$ a *feasible point*.

We write the constrained optimization problem as

$$\begin{aligned} &\min_b f(b) = (y - Xb)^{\mathrm{T}}(y - Xb) \\ &\text{s.t. } Ab = c. \end{aligned} \tag{4.25}$$

If b_c is any feasible point (that is, $Ab_c = c$), then any other feasible point can be represented as $b_c + p$, where p is any vector in the null space of A, $\mathcal{N}(A)$. From our discussion in Section 3.5.2, we know that the dimension of $\mathcal{N}(A)$ is $m-k$, and its order is m. If N is an $m\times(m-k)$ matrix whose columns form a basis for $\mathcal{N}(A)$, all feasible points can be generated by $b_c + Nz$, where $z \in \mathrm{I\!R}^{m-k}$. Hence, we need only consider the restricted variables

$$b = b_c + Nz$$

and the "reduced" function

$$h(z) = f(b_c + Nz).$$

The argument of this function is a vector with only $m - k$ elements instead of m elements as in the unconstrained problem. The unconstrained minimum of h, however, is the solution of the original constrained problem.

The Reduced Gradient and Reduced Hessian

If we assume differentiability, the gradient and Hessian of the reduced function can be expressed in terms of the original function:

$$\begin{aligned} \mathrm{g}_h(z) &= N^{\mathrm{T}}\mathrm{g}_f(b_c + Nz) \\ &= N^{\mathrm{T}}\mathrm{g}_f(b) \end{aligned} \tag{4.26}$$

and

$$\begin{aligned} \mathrm{H}_h(z) &= N^{\mathrm{T}}\mathrm{H}_f(b_c + Nz)N \\ &= N^{\mathrm{T}}\mathrm{H}_f(b)N. \end{aligned} \tag{4.27}$$

In equation (4.26), $N^{\mathrm{T}}\mathrm{g}_f(b)$ is called the *reduced gradient* or *projected gradient*, and $N^{\mathrm{T}}\mathrm{H}_f(b)N$ in equation (4.27) is called the *reduced Hessian* or *projected Hessian*.

The properties of stationary points are related to the derivatives referred to above are the conditions that determine a minimum of this reduced objective function; that is, b_* is a minimum if and only if

- $N^{\mathrm{T}}\mathrm{g}_f(b_*) = 0$,
- $N^{\mathrm{T}}\mathrm{H}_f(b_*)N$ is positive definite, and
- $Ab_* = c$.

These relationships then provide the basis for the solution of the optimization problem.

Lagrange Multipliers

Because the $m \times m$ matrix $[N|A^{\mathrm{T}}]$ spans $\mathrm{I\!R}^m$, we can represent the vector $\mathrm{g}_f(b_*)$ as a linear combination of the columns of N and A^{T}, that is,

$$\begin{aligned} \mathrm{g}_f(b_*) &= [N|A^{\mathrm{T}}] \begin{pmatrix} z_* \\ \lambda_* \end{pmatrix} \\ &= \begin{pmatrix} Nz_* \\ A^{\mathrm{T}}\lambda_* \end{pmatrix}, \end{aligned}$$

where z_* is an $(m-k)$-vector and λ_* is a k-vector. Because $\nabla h(z_*) = 0$, Nz_* must also vanish (that is, $Nz_* = 0$), and thus, at the optimum, the nonzero elements of the gradient of the objective function are linear combinations of the rows of the constraint matrix, $A^{\mathrm{T}}\lambda_*$. The k elements of the linear combination vector λ_* are called *Lagrange multipliers*.

The Lagrangian

Let us now consider a simple generalization of the constrained problem above and an abstraction of the results above so as to develop a general method. We consider the problem

$$\begin{aligned} &\min_x \; f(x) \\ &\text{s.t.}\; c(x) = 0, \end{aligned} \tag{4.28}$$

where f is a scalar-valued function of an m-vector variable and c is a k-vector-valued function of the variable. There are some issues concerning the equation $c(x) = 0$ that we will not go into here. Obviously, we have the same concerns as before; that is, whether $c(x) = 0$ is consistent and whether the individual equations $c_i(x) = 0$ are independent. Let us just assume they are and proceed. (Again, we refer the interested reader to a more general text on optimization.)

Motivated by the results above, we form a function that incorporates a dot product of Lagrange multipliers and the function $c(x)$:

$$F(x) = f(x) + \lambda^{\mathrm{T}} c(x). \tag{4.29}$$

This function is called the *Lagrangian.* The solution, (x_*, λ_*), of the optimization problem occurs at a stationary point of the Lagrangian,

$$\mathrm{g}_f(x_*) = \begin{pmatrix} 0 \\ \mathrm{J}_c(x_*)^{\mathrm{T}} \lambda_* \end{pmatrix}. \tag{4.30}$$

Thus, at the optimum, the gradient of the objective function is a linear combination of the columns of the Jacobian of the constraints.

Another Example: The Rayleigh Quotient

The important equation (3.208) on page 122 can also be derived by using differentiation. This equation involves maximization of the Rayleigh quotient (equation (3.209)),

$$x^{\mathrm{T}} A x / x^{\mathrm{T}} x$$

under the constraint that $x \neq 0$. In this function, this constraint is equivalent to the constraint that $x^{\mathrm{T}} x$ equal a fixed nonzero constant, which is canceled in the numerator and denominator. We can arbitrarily require that $x^{\mathrm{T}} x = 1$, and the problem is now to determine the maximum of $x^{\mathrm{T}} A x$ subject to the constraint $x^{\mathrm{T}} x = 1$. We now formulate the Lagrangian

$$x^{\mathrm{T}} A x - \lambda(x^{\mathrm{T}} x - 1), \tag{4.31}$$

differentiate, and set it equal to 0, yielding

$$Ax - \lambda x = 0.$$

This implies that a stationary point of the Lagrangian occurs at an eigenvector and that the value of $x^{\mathrm{T}}Ax$ is an eigenvalue. This leads to the conclusion that the maximum of the ratio is the maximum eigenvalue. We also see that the second order necessary condition for a local maximum is satisfied; $A - \lambda I$ is nonpositive definite when λ is the maximum eigenvalue. (We can see this using the spectral decomposition of A and then subtracting λI.) Note that we do not have the sufficient condition that $A - \lambda I$ is negative definite ($A - \lambda I$ is obviously singular), but the fact that it is a maximum is established by inspection of the finite set of stationary points.

Optimization without Differentiation

In the previous example, differentiation led us to a stationary point, but we had to establish by inspection that the stationary point is a maximum. In optimization problems generally, and in constrained optimization problems particularly, it is often easier to use other methods to determine the optimum.

A constrained minimization problem we encounter occasionally is

$$\min_X \left(\log|X| + \mathrm{tr}(X^{-1}A)\right) \tag{4.32}$$

for a given positive definite matrix A and subject to X being positive definite. The derivatives given in Table 4.2 could be used. The derivatives set equal to 0 immediately yield $X = A$. This means that $X = A$ is a stationary point, but whether or not it is a minimum would require further analysis. As is often the case with such problems, an alternate approach leaves no such pesky complications. Let A and X be $n \times n$ positive definite matrices, and let $c_1, \ldots, c_n$ be the eigenvalues of $X^{-1}A$. Now, by property 7 on page 107 these are also the eigenvalues of $X^{-1/2}AX^{-1/2}$, which is positive definite (see inequality (3.122) on page 89). Now, consider the expression (4.32) with general X minus the expression with $X = A$:

$$\begin{aligned}
\log|X| + \mathrm{tr}(X^{-1}A) - \log|A| - \mathrm{tr}(A^{-1}A) &= \log|XA^{-1}| + \mathrm{tr}(X^{-1}A) - \mathrm{tr}(I) \\
&= -\log|X^{-1}A| + \mathrm{tr}(X^{-1}A) - n \\
&= -\log\left(\prod_i c_i\right) + \sum_i c_i - n \\
&= \sum_i (-\log c_i + c_i - 1) \\
&\geq 0
\end{aligned}$$

because if $c > 0$, then $\log c \leq c - 1$, and the minimun occurs when each $c_i = 1$; that is, when $X^{-1}A = I$. Thus, the minimum of expression (4.32) occurs uniquely at $X = A$.

4.4 Multiparameter Likelihood Functions

For a sample $y = (y_1, \ldots, y_n)$ from a probability distribution with probability density function $p(\cdot;\, \theta)$, the *likelihood function* is

$$L(\theta;\, y) = \prod_{i=1}^{n} p(y_i;\, \theta), \tag{4.33}$$

and the *log-likelihood function* is $l(\theta;\, y) = \log(L(\theta;\, y))$. It is often easier to work with the log-likelihood function.

The log-likelihood is an important quantity in information theory and in unbiased estimation. If Y is a random variable with the given probability density function with the r-vector parameter θ, the *Fisher information* matrix that Y contains about θ is the $r \times r$ matrix

$$I(\theta) = \mathrm{Cov}_\theta \left(\frac{\partial l(t, Y)}{\partial t_i}, \frac{\partial l(t, Y)}{\partial t_j} \right), \tag{4.34}$$

where Cov_θ represents the variance-covariance matrix of the functions of Y formed by taking expectations for the given θ. (I use different symbols here because the derivatives are taken with respect to a *variable*, but the θ in Cov_θ cannot be the variable of the differentiation. This distinction is somewhat pedantic, and sometimes I follow the more common practice of using the same symbol in an expression that involves both Cov_θ and $\partial l(\theta, Y)/\partial\theta_i$.)

For example, if the distribution is the d-variate normal distribution with mean d-vector μ and $d \times d$ positive definite variance-covariance matrix Σ, the likelihood, equation (4.33), is

$$L(\mu, \Sigma;\, y) = \frac{1}{\left((2\pi)^{d/2} |\Sigma|^{1/2}\right)^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^{\mathrm{T}} \Sigma^{-1} (y_i - \mu) \right).$$

(Note that $|\Sigma|^{1/2} = |\Sigma^{\frac{1}{2}}|$. The square root matrix $\Sigma^{\frac{1}{2}}$ is often useful in transformations of variables.)

Anytime we have a quadratic form that we need to simplify, we should recall equation (3.63): $x^{\mathrm{T}} A x = \mathrm{tr}(A x x^{\mathrm{T}})$. Using this, and because, as is often the case, the log-likelihood is easier to work with, we write

$$l(\mu, \Sigma;\, y) = c - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)^{\mathrm{T}} \right), \tag{4.35}$$

where we have used c to represent the constant portion. Next, we use the Pythagorean equation (2.47) or equation (3.64) on the outer product to get

$$\begin{aligned} l(\mu, \Sigma;\, y) = c - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}} \right) \\ - \frac{n}{2} \mathrm{tr}\left(\Sigma^{-1} (\bar{y} - \mu)(\bar{y} - \mu)^{\mathrm{T}} \right). \end{aligned} \tag{4.36}$$

In maximum likelihood estimation, we seek the maximum of the likelihood function (4.33) with respect to θ while we consider y to be fixed. If the maximum occurs within an open set and if the likelihood is differentiable, we might be able to find the maximum likelihood estimates by differentiation. In the log-likelihood for the d-variate normal distribution, we consider the parameters μ and Σ to be variables. To emphasize that perspective, we replace the parameters μ and Σ by the variables $\hat{\mu}$ and $\widehat{\Sigma}$. Now, to determine the maximum, we could take derivatives with respect to $\hat{\mu}$ and $\widehat{\Sigma}$, set them equal to 0, and solve for the maximum likelihood estimates. Some subtle problems arise that depend on the fact that for any constant vector a and scalar b, $\Pr(a^{\mathrm{T}}X = b) = 0$, but we do not interpret the likelihood as a probability. In Exercise 4.5b you are asked to determine the values of $\hat{\mu}$ and $\widehat{\Sigma}$ using properties of traces and positive definite matrices without resorting to differentiation. (This approach does not avoid the subtle problems, however.)

Often in working out maximum likelihood estimates, students immediately think of differentiating, setting to 0, and solving. As noted above, this requires that the likelihood function be differentiable, that it be concave, and that the maximum occur at an interior point of the parameter space. Keeping in mind exactly what the problem is — one of finding a maximum — often leads to the correct solution more quickly.

4.5 Integration and Expectation

Just as we can take derivatives with respect to vectors or matrices, we can also take antiderivatives or definite integrals with respect to vectors or matrices. Our interest is in integration of functions weighted by a multivariate probability density function, and for our purposes we will be interested only in definite integrals.

Again, there are three components:

- the differential (the variable of the operation) and its domain (the range of the integration),
- the integrand (the function), and
- the result of the operation (the integral).

In the simplest case, all three of these objects are of the same type; they are scalars. In the happy cases that we consider, each definite integral within the nested sequence exists, so convergence and order of integration are not issues. (The implication of these remarks is that while there is a much bigger field of mathematics here, we are concerned about the relatively simple cases that suffice for our purposes.)

In some cases of interest involving vector-valued random variables, the differential is the vector representing the values of the random variable and the integrand has a scalar function (the probability density) as a factor. In one type of such an integral, the integrand is only the probability density function, and the integral evaluates to a probability, which of course is a scalar. In another type of such an integral, the integrand is a vector representing the values of the random variable times the probability density function. The integral in this case evaluates to a vector, namely the expectation of the random variable over the domain of the integration. Finally, in an example of a third type of such an integral, the integrand is an outer product with itself of a vector representing the values of the random variable minus its mean times the probability density function. The integral in this case evaluates to a variance-covariance matrix. In each of these cases, the integral is the same type of object as the integrand.

4.5.1 Multidimensional Integrals and Integrals Involving Vectors and Matrices

An integral of the form $\int f(v)\,\mathrm{d}v$, where v is a vector, can usually be evaluated as a multiple integral with respect to each differential $\mathrm{d}v_i$. Likewise, an integral of the form $\int f(M)\,\mathrm{d}M$, where M is a matrix can usually be evaluated by "unstacking" the columns of $\mathrm{d}M$, evaluating the integral as a multiple integral with respect to each differential $\mathrm{d}m_{ij}$, and then possibly "restacking" the result.

Multivariate integrals (that is, integrals taken with respect to a vector or a matrix) define probabilities and expectations in multivariate probability distributions. As with many well-known univariate integrals, such as $\Gamma(\cdot)$, that relate to univariate probability distributions, there are standard multivariate integrals, such as the multivariate gamma, $\Gamma_d(\cdot)$, that relate to multivariate probability distributions. Using standard integrals often facilitates the computations.

Change of Variables; Jacobians

When evaluating an integral of the form $\int f(x)\,\mathrm{d}x$, where x is a vector, for various reasons we may form a one-to-one differentiable transformation of the variables of integration; that is, of x. We write x as a function of the new variables; that is, $x = g(y)$, and so $y = g^{-1}(x)$. A simple fact from elementary multivariable calculus is

$$\int_{R(x)} f(x)\,\mathrm{d}x = \int_{R(y)} f(g(y))\,|\mathrm{det}(\mathrm{J}_g(y))|\mathrm{d}y, \tag{4.37}$$

where $R(y)$ is the image of $R(x)$ under g^{-1} and $\mathrm{J}_g(y)$ is the Jacobian of g (see equation (4.12)). (This is essentially a chain rule result for $\mathrm{d}x = \mathrm{d}(g(y)) = \mathrm{J}_g\mathrm{d}y$ under the interpretation of $\mathrm{d}x$ and $\mathrm{d}y$ as positive differential elements and the interpretation of $|\mathrm{det}(\mathrm{J}_g)|$ as a volume element, as discussed on page 57.)

In the simple case of a full rank linear transformation of a vector, the Jacobian is constant, and so for $y = Ax$ with A a fixed matrix, we have

$$\int f(x)\,\mathrm{d}x = |\det(A)|^{-1} \int f(A^{-1}y)\,\mathrm{d}y.$$

(Note that we write $\det(A)$ instead of $|A|$ for the determinant if we are to take the absolute value of it because otherwise we would have $||A||$, which is a symbol for a norm. However, $|\det(A)|$ is not a norm; it lacks each of the properties listed on page 16.)

In the case of a full rank linear transformation of a matrix variable of integration, the Jacobian is somewhat more complicated, but the Jacobian is constant for a fixed transformation matrix. For a transformation $Y = AX$, we determine the Jacobian as above by considering the columns of X one by one. Hence, if X is an $n \times m$ matrix and A is a constant nonsingular matrix, we have

$$\int f(X)\,\mathrm{d}X = |\det(A)|^{-m} \int f(A^{-1}Y)\,\mathrm{d}Y.$$

For a transformation of the form $Z = XB$, we determine the Jacobian by considering the rows of X one by one.

4.5.2 Integration Combined with Other Operations

Integration and another finite linear operator can generally be performed in any order. For example, because the trace is a finite linear operator, integration and the trace can be performed in either order:

$$\int \mathrm{tr}(A(x))\mathrm{d}x = \mathrm{tr}\left(\int A(x)\mathrm{d}x\right).$$

For a scalar function of two vectors x and y, it is often of interest to perform differentiation with respect to one vector and integration with respect to the other vector. In such cases, it is of interest to know when these operations can be interchanged. The answer is given in the following theorem, which is a consequence of the Lebesgue dominated convergence theorem. Its proof can be found in any standard text on real analysis.

> Let $\mathcal{X}$ be an open set, and let $f(x,y)$ and $\partial f/\partial x$ be scalar-valued functions that are continuous on $\mathcal{X} \times \mathcal{Y}$ for some set $\mathcal{Y}$. Now suppose there are scalar functions $g_0(y)$ and $g_1(y)$ such that
>
> $$\left.\begin{array}{r} |f(x,y)| \le g_0(y) \\ \\ \|\frac{\partial}{\partial x} f(x,y)\| \le g_1(y) \end{array}\right\} \quad \text{for all } (x,y) \in \mathcal{X} \times \mathcal{Y},$$
>
> $$\int_{\mathcal{Y}} g_0(y)\,\mathrm{d}y < \infty,$$
>
> and
>
> $$\int_{\mathcal{Y}} g_1(y)\,\mathrm{d}y < \infty.$$

Then

$$\frac{\partial}{\partial x}\int_{\mathcal{Y}} f(x,y)\,\mathrm{d}y = \int_{\mathcal{Y}} \frac{\partial}{\partial x} f(x,y)\,\mathrm{d}y. \tag{4.38}$$

An important application of this interchange is in developing the information inequality. (This inequality is not germane to the present discussion; it is only noted here for readers who may already be familiar with it.)

4.5.3 Random Variables

A vector random variable is a function from some sample space into $\mathbb{R}^n$, and a matrix random variable is a function from a sample space into $\mathbb{R}^{n\times m}$. (Technically, in each case, the function is required to be *measurable* with respect to a *measure* defined in the context of the sample space and an appropriate collection of subsets of the sample space.) Associated with each random variable is a distribution function whose derivative with respect to an appropriate measure is nonnegative and integrates to 1 over the full space formed by $\mathbb{R}$.

Vector Random Variables

The simplest kind of vector random variable is one whose elements are independent. Such random vectors are easy to work with because the elements can be dealt with individually, but they have limited applications. More interesting random vectors have a multivariate structure that depends on the relationships of the distributions of the individual elements. The simplest nondegenerate multivariate structure is of second degree; that is, a covariance or correlation structure. The probability density of a random vector with a multivariate structure generally is best represented by using matrices. In the case of the multivariate normal distribution, the variances and covariances together with the means completely characterize the distribution. For example, the fundamental integral that is associated with the d-variate normal distribution, sometimes called *Aitken's integral*,

$$\int_{\mathbb{R}^d} \mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}\,\mathrm{d}x = (2\pi)^{d/2}|\Sigma|^{1/2}, \tag{4.39}$$

provides that constant. The rank of the integral is the same as the rank of the integrand. ("Rank" is used here in the sense of "number of dimensions".) In this case, the integrand and the integral are scalars.

Equation (4.39) is a simple result that follows from the evaluation of the individual single integrals after making the change of variables $y_i = x_i - \mu_i$. It can also be seen by first noting that because Σ^{-1} is positive definite, as in equation (3.215), it can be written as $P^{\mathrm{T}}\Sigma^{-1}P = I$ for some nonsingular matrix P. Now, after the translation $y = x - \mu$, which leaves the integral unchanged, we make the linear change of variables $z = P^{-1}y$, with the associated Jacobian $|\det(P)|$, as in equation (4.37). From $P^{\mathrm{T}}\Sigma^{-1}P = I$, we have

$|\det(P)| = (\det(\Sigma))^{1/2} = |\Sigma|^{1/2}$ because the determinant is positive. Aitken's integral therefore is

$$\begin{aligned}\int_{\mathrm{I\!R}^d} \mathrm{e}^{-y^{\mathrm{T}}\Sigma^{-1}y/2}\,\mathrm{d}y &= \int_{\mathrm{I\!R}^d} \mathrm{e}^{-(Pz)^{\mathrm{T}}\Sigma^{-1}Pz/2}\,(\det(\Sigma))^{1/2}\mathrm{d}z \\ &= \int_{\mathrm{I\!R}^d} \mathrm{e}^{-z^{\mathrm{T}}z/2}\,\mathrm{d}z\,(\det(\Sigma))^{1/2} \\ &= (2\pi)^{d/2}(\det(\Sigma))^{1/2}.\end{aligned}$$

The expected value of a function f of the vector-valued random variable X is

$$\mathrm{E}(f(X)) = \int_{D(X)} f(x)p_X(x)\,\mathrm{d}x, \qquad (4.40)$$

where $D(X)$ is the support of the distribution, $p_X(x)$ is the probability density function evaluated at x, and x dx are dummy vectors whose elements correspond to those of X. Interpreting $\int_{D(X)}\mathrm{d}x$ as a nest of univariate integrals, the result of the integration of the vector $f(x)p_X(x)$ is clearly of the same type as $f(x)$. For example, if $f(x) = x$, the expectation is the mean, which is a vector. For the normal distribution, we have

$$\begin{aligned}\mathrm{E}(X) &= (2\pi)^{-d/2}|\Sigma|^{-1/2}\int_{\mathrm{I\!R}^d} x\mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}\,\mathrm{d}x \\ &= \mu.\end{aligned}$$

For the variance of the vector-valued random variable X,

$$\mathrm{V}(X),$$

the function f in expression (4.40) above is the matrix $(X - \mathrm{E}(X))(X - \mathrm{E}(X))^{\mathrm{T}}$, and the result is a matrix. An example is the normal variance:

$$\begin{aligned}\mathrm{V}(X) &= \mathrm{E}\left((X-\mathrm{E}(X))(X-\mathrm{E}(X))^{\mathrm{T}}\right) \\ &= (2\pi)^{-d/2}|\Sigma|^{-1/2}\int_{\mathrm{I\!R}^d}\left((x-\mu)(x-\mu)^{\mathrm{T}}\right)\mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}\,\mathrm{d}x \\ &= \Sigma.\end{aligned}$$

Matrix Random Variables

While there are many random variables of interest that are vectors, there are only a few random matrices whose distributions have been studied. One, of course, is the Wishart distribution; see Exercise 4.8. An integral of the Wishart probability density function over a set of nonnegative definite matrices is the probability of the set.

A simple distribution for random matrices is one in which the individual elements have identical and independent normal distributions. This distribution of matrices was named the BMvN distribution by Birkhoff and Gulati (1979) (from the last names of three mathematicians who used such random matrices in numerical studies). Birkhoff and Gulati (1979) showed that if the elements of the $n \times n$ matrix X are i.i.d. $\mathrm{N}(0, \sigma^2)$, and if Q is an orthogonal matrix and R is an upper triangular matrix with positive elements on the diagonal such that $QR = X$, then Q has the *Haar distribution*. (The factorization $X = QR$ is called the QR decomposition and is discussed on page 190 If X is a random matrix as described, this factorization exists with probability 1.) The Haar(n) distribution is uniform over the space of $n \times n$ orthogonal matrices.

The measure

$$\mu(D) = \int_D H^{\mathrm{T}} \,\mathrm{d}H, \tag{4.41}$$

where D is a subset of the orthogonal group $\mathcal{O}(n)$ (see page 105), is called the *Haar measure*. This measure is used to define a kind of "uniform" probability distribution for orthogonal factors of random matrices. For any $Q \in \mathcal{O}(n)$, let QD represent the subset of $\mathcal{O}(n)$ consisting of the matrices $\widetilde{H} = QH$ for $H \in D$ and DQ represent the subset of matrices formed as HQ. From the integral, we see

$$\mu(QD) = \mu(DQ) = \mu(D),$$

so the Haar measure is invariant to multiplication within the group. The measure is therefore also called the *Haar invariant measure* over the orthogonal group. (See Muirhead, 1982, for more properties of this measure.)

A common matrix integral is the complete d-variate gamma function, denoted by $\Gamma_d(x)$ and defined as

$$\Gamma_d(x) = \int_D \mathrm{e}^{-\mathrm{tr}(A)} |A|^{x-(d+1)/2} \,\mathrm{d}A, \tag{4.42}$$

where D is the set of all $d \times d$ positive definite matrices, $A \in D$, and $x > (d-1)/2$. A multivariate gamma distribution can be defined in terms of the integrand. (There are different definitions of a multivariate gamma distribution.) The multivariate gamma function also appears in the probability density function for a Wishart random variable (see Muirhead, 1982, or Carmeli, 1983, for example).

Exercises

4.1. Use equation (4.6), which defines the derivative of a matrix with respect to a scalar, to show the product rule equation (4.3) directly:

$$\frac{\partial YW}{\partial x} = \frac{\partial Y}{\partial x} W + Y \frac{\partial W}{\partial x}.$$

4.2. For the n-vector x, compute the gradient $\mathrm{g_V}(x)$, where $\mathrm{V}(x)$ is the variance of x, as given in equation (2.53).

Hint: Use the chain rule.

4.3. For the square, nonsingular matrix Y, show that

$$\frac{\partial Y^{-1}}{\partial x} = -Y^{-1}\frac{\partial Y}{\partial x}Y^{-1}.$$

Hint: Differentiate $YY^{-1} = I$.

4.4. Newton's method.

You should not, of course, just blindly pick a starting point and begin iterating. How can you be sure that your solution is a local optimum? Can you be sure that your solution is a global optimum? It is often a good idea to make some plots of the function. In the case of a function of a single variable, you may want to make plots in different scales. For functions of more than one variable, profile plots may be useful (that is, plots of the function in one variable with all the other variables held constant).

a) Use Newton's method to determine the maximum of the function $f(x) = \sin(4x) - x^4/12$.

b) Use Newton's method to determine the minimum of

$$f(x_1, x_2) = 2x_1^4 + 3x_1^3 + 2x_1^2 + x_2^2 - 4x_1x_2.$$

What is the Hessian at the minimum?

4.5. Consider the log-likelihood $l(\mu, \Sigma; y)$ for the d-variate normal distribution, equation (4.35). Be aware of the subtle issue referred to in the text. It has to do with whether $\sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}$ is positive definite.

a) Replace the parameters μ and Σ by the variables $\hat{\mu}$ and $\widehat{\Sigma}$, take derivatives with respect to $\hat{\mu}$ and $\widehat{\Sigma}$, set them equal to 0, and solve for the maximum likelihood estimates. What assumptions do you have to make about n and d?

b) Another approach to maximizing the expression in equation (4.35) is to maximize the last term with respect to $\hat{\mu}$ (this is the only term involving μ) and then, with the maximizing value substituted, to maximize

$$-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}\right).$$

Use this approach to determine the maximum likelihood estimates $\hat{\mu}$ and $\widehat{\Sigma}$.

4.6. Let

$$D = \left\{\begin{bmatrix} c & -s \\ s & c \end{bmatrix} : -1 \leq c \leq 1, c^2 + s^2 = 1\right\}.$$

Evaluate the Haar measure $\mu(D)$. (This is the class of 2 × 2 rotation matrices; see equation (5.3), page 177.)

4.7. Write a Fortran or C program to generate $n \times n$ random orthogonal matrices with the Haar uniform distribution. Use the following method due to Heiberger (1978), which was modified by Stewart (1980). (See also Tanner and Thisted, 1982.)

a) Generate $n-1$ independent i-vectors, $x_2, x_3, \ldots, x_n$, from $N_i(0, I_i)$. (x_i is of length i.)

b) Let $r_i = \|x_i\|_2$, and let $\widetilde{H}_i$ be the $i \times i$ reflection matrix that transforms x_i into the i-vector $(r_i, 0, 0, \ldots, 0)$.

c) Let H_i be the $n \times n$ matrix

$$\begin{bmatrix} I_{n-i} & 0 \\ 0 & \widetilde{H}_i \end{bmatrix},$$

and form the diagonal matrix,

$$J = \mathrm{diag}\big((-1)^{b_1}, (-1)^{b_2}, \ldots, (-1)^{b_n}\big),$$

where the b_i are independent realizations of a Bernoulli random variable.

d) Deliver the orthogonal matrix $Q = JH_1H_2 \cdots H_n$.

The matrix Q generated in this way is orthogonal and has a Haar distribution.

Can you think of any way to test the goodness-of-fit of samples from this algorithm? Generate a sample of 1,000 2×2 random orthogonal matrices, and assess how well the sample follows a Haar uniform distribution.

4.8. The probability density for the Wishart distribution is proportional to

$$\mathrm{e}^{\mathrm{tr}(\Sigma^{-1}W/2)}|W|^{(n-d-1)/2},$$

where W is a $d \times d$ nonnegative definite matrix, the parameter Σ is a fixed $d \times d$ positive definite matrix, and the parameter n is positive. (Often n is restricted to integer values greater than d.) Determine the constant of proportionality.