Multivariate Maxima and Minima with Matrix Derivatives

Derrick S. Tracy; Paul S. Dwyer


Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196912%2964%3A328%3C1576%3AMAMW%3E2.0.CO%3B2-B

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

*Journal of the American Statistical Association* is published by American Statistical Association. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

©1969 American Statistical Association

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2003 JSTOR
MULTIVARIATE MAXIMA AND MINIMA
WITH MATRIX DERIVATIVES

DERRICK S. TRACY*
University of Windsor

and

PAUL S. DWYER
University of Michigan

The purpose of this paper is the presentation of formulae for obtaining
matrix derivatives of the second order to use in making tests for maxima
and minima. The theory of such second order derivatives is presented.
These formulae require the rearrangement of the parameter elements in
vector form and the transformed results feature Kronecker products,
which have certain desirable properties. Application is made to several
types of problems.

1. INTRODUCTION AND SUMMARY

This paper is a sequel to the paper on the use of matrix derivatives in multi-
variate analysis [6] in which it is shown that appropriate first order
matrix derivatives are easily obtained and are applicable to multivariate prob-
lems including the problem of extrema. The chief objective of this paper is to
carry the process one step further and to show how formulae for second order
matrix derivatives, appropriate for use in testing maxima and minima, can be
obtained and applied.

The case of vector derivatives of vectors, a special case of the problem of
matrix derivatives, is discussed in Section 2. Here, in \( \frac{\partial y}{\partial x} \), the vector deriva-
tive of vector \( y \) with respect to vector \( x \), the elements of \( y \) appear as a row and
those of \( x \) as a column.

As in the earlier paper [6], the matrix derivative is taken with respect to the
matrix element, identified by the three components consisting of a) the row,
b) the column, and c) the scalar value. There are no equality relations possible
between such matrix elements even though relations, such as those of symmetry,
may exist between the scalar values. As a result, any relations between the
scalar values may be ignored in the general differentiation process (matrix ele-
ments) though they may be used, once the differentiation is complete, in ob-
taining the specific result.

Arrangements are suggested for collecting the elements of matrices in vector
form. The elements of matrix \( Y \) are displayed as a row vector \( Y \), and those of
matrix \( X \) as a column vector \( X \), and the formulae for \( \frac{\partial Y}{\partial X} \), are obtained.
Other vector arrangements of the elements of \( Y \) and \( X \) are examined in Section
3. Kronecker products of matrices play an important role in this development.
Some of their relevant properties are presented in Appendix A.

The theory of second order derivatives useful in testing extrema is discussed
in Section 4. Necessary and sufficient conditions, resting on the positive or
negative definiteness of the Hessian matrix at the critical value, are stated.

* Research supported (in part) by the Defence Research Board of Canada, Grant Number 9540–15.

1576
A theorem which is useful when extremizing a convex or concave function is indicated. The case of constrained extrema is discussed. Some forms of Hessian matrices which occur commonly in applications are discussed in Appendix B, while Appendix C gives some rules for determining constrained extrema.

These results are applied in Section 5 to the problem of testing for extrema. The first set of illustrations deals with (i) the optimization of the ratio of two quadratic forms, and (ii) the optimization of a quadratic form subject to linear constraints. The second set of illustrations deals with least squares and general least squares theory. The third set of illustrations treats the problem of testing joint maximum likelihood estimates of the parameters (i) of the multivariate normal distribution, (ii) of multivariate normal regression, and (iii) of the general linear hypothesis using $K$ subgroups with missing observations.

2. VECTOR DERIVATIVES OF VECTORS

If $x$ and $y$ are vectors,\footnote{Throughout the paper, a vector is treated as a special case of a matrix, which consists of certain scalar elements in specified positions.} with general elements $x_i$, $y_j$, we can define the vector derivative $\frac{\partial y}{\partial x}$ to be the matrix $(\partial y_j/\partial x_i)$. In this arrangement of derivatives $\partial y_j/\partial x_i$, the elements of $y$ appear as a row and those of $x$ as a column. If $x$ or $y$ or both are scalars, the derivative is a row vector or a column vector or a scalar respectively.

Consider $y = Ax$, where $x$, $y$ are both column vectors and $A$ the matrix $(a_{ij})$. Then $y_i = \sum_j a_{ij}x_j$ yields

$$\frac{\partial y}{\partial x} = (a_{ii}) = A^T.$$

If one had $y = x^TA$, where $y$ is a row vector and $x$ a column vector,

$$\frac{\partial y}{\partial x} = (a_{ij}) = A.$$

If $y$ is a scalar function of a vector $x$, then $\partial y/\partial x$ is a vector, so $\partial/\partial x[\partial y/\partial x]$ is the derivative of a vector with respect to a vector and the above rules apply.

Thus these results are adequate for testing the extrema of many statistical functions which are scalar functions of vectors. For example, the (symmetric) quadratic form $y = x^TAx$ yields

$$\frac{\partial y}{\partial x} = 2Ax$$

[1, p. 347], [6, p. 609], [11, p. 12], [15, p. 57] and hence

$$\frac{\partial}{\partial x} \frac{\partial y}{\partial x} = 2A.$$

Some vector derivative formulae with $y$ a vector, quadratic form or bilinear form are presented in Table 1.
TABLE 1. FORMULAE FOR SOME VECTOR DERIVATIVES

\[
\begin{array}{c|c|c}
\text{\(x, z\) column vectors} & \frac{\partial y}{\partial x} & (2.1) \\
\hline
A^x & A^\tau & (2.1') \\
x^\tau A & A & (2.2) \\
x^\tau x & 2x & (2.3) \\
X^\tau A x & 2Ax & (2.4) \\
x^\tau A z & Az & (2.5) \\
\end{array}
\]

\[
\begin{array}{c|c|c}
\text{\(x, z\) row vectors} & \frac{\partial y}{\partial x} & (2.1') \\
\hline
A^x & A^\tau & (2.1') \\
x A & A & (2.2') \\
x x^\tau & 2x^\tau & (2.3') \\
x A x^\tau & 2 A x^\tau & (2.4') \\
x A z^\tau & A z^\tau & (2.5') \\
\end{array}
\]

3. ARRANGEMENTS OF MATRIX DERIVATIVE ELEMENTS

If \(X, Y\) are two matrices, let \(\langle X \rangle_{\alpha\beta}\) denote the \(\alpha, \beta\) element of \(X\) as in [6], and similarly \(\langle Y \rangle_{\gamma\delta}\). Less formally, \(\langle X \rangle, \langle Y \rangle\) are used to denote these elements. \(J_{\alpha\beta}\), and less formally \(J\), is used for a matrix having the same dimensions as \(X\), with 1 in the \(\alpha, \beta\) position and 0's elsewhere. \(K\) is correspondingly used for a matrix like \(Y\) with 1 in the \(\gamma, \delta\) position. Clearly

\[
\frac{\partial X}{\partial \langle X \rangle_{\alpha\beta}} = J_{\alpha\beta}^2 \quad \text{and} \quad \frac{\partial \langle X \rangle_{\gamma\delta}}{\partial X} = K_{\gamma\delta}.
\]

(When dealing with vectors \(x, y\), we use vectors \(j, k\).)

When \(Y = A X B\),

\[
\frac{\partial Y}{\partial \langle X \rangle_{\alpha\beta}} = AJ_{\alpha\beta}B.
\]

No matter whether \(A\) and \(B\) are functions of \(X\) or not, it is shown in [6, pp. 611–612] that with

\[
\frac{\partial Y}{\partial \langle X \rangle} = AJB
\]

one has

\[
\frac{\partial \langle Y \rangle}{\partial X} = A^\tau KB^\tau \quad \text{and} \quad \frac{\partial \langle Y \rangle_{\gamma\delta}}{\partial \langle X \rangle_{\alpha\beta}} = a_{\gamma\alpha} b_{\beta\delta}.
\]

A way to collect these derivatives is suggested by rearranging the elements of the matrices in vector form. The matrix elements of \(Y\) are placed in a row vector and those of \(X\) in a column vector. We denote by \(X_\tau/X_e\) the vector presentation of all elements of \(X\) in which the elements of the first row/column appear in order first, then the elements of the second row/column and so on, and similarly for \(Y\). Then

---

1 This is the matrix derivative element in which the derivative is taken with respect to the matrix element (the scalar value in a specified matrix position) and not with respect to the scalar value of the element [6, p. 608].

2 This material is presented here in very concise form for purposes of this paper. A more adequate presentation is available in [6].

3 The idea of \(X_e\) goes back as far as Koopmans [12] and has been used by econometricians, e.g. Rothenberg and Leenders [16], Neudecker [14]. They use vec \(X\) for our \(X_e\).
\[
\frac{\partial Y_r}{\partial X_r} = A^T \otimes B,
\]

the direct or Kronecker product of \(A^T\) and \(B\), since this arrangement of terms \(a_{\gamma \alpha}b_{\beta \tau}\) presents all elements with the same \(a_{\gamma \alpha}\) as a block \(a_{\gamma \alpha}B\), with \(\gamma > \alpha\) in blocks in the upper triangular part of the matrix, and the collection of all these blocks is \(A^T \otimes B\). Similarly

\[
\frac{\partial Y_e}{\partial X_e} = B \otimes A^T.
\]

If we

(i) let \(X\) be \(m \times n\),
(ii) let matrices \(A, B, C, D\) be functions of \(X\) or not,
(iii) let \(M^{(k)}\) represent a matrix obtained by rearranging the rows of matrix \(M\) by taking every \(k\)th row starting with the first, then every \(k\)th row starting with the second, etc., \(^5\)
(iv) notice that when

\[
\frac{\partial Y}{\partial \langle X \rangle} = CJ^T D,
\]

we can substitute \(Z^T\) for \(X\), obtaining

\[
\frac{\partial Y}{\partial \langle Z^T \rangle} = CJ^T D \quad \text{or} \quad \frac{\partial Y}{\partial \langle Z \rangle} = CJ D,
\]

so

\[
\frac{\partial Y_r}{\partial Z_r} = \frac{\partial Y_r}{\partial X_e} = C^T \otimes D,
\]

then for

\[
\frac{\partial Y}{\partial \langle X \rangle} = AJB
\]

or \(CJ^T D\), we get entries in Table 2 as formulae for

**TABLE 2. EXPRESSIONS FOR**

<table>
<thead>
<tr>
<th>Partial of (\rightarrow)</th>
<th>(Y_r)</th>
<th>(Y_e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\frac{\partial Y}{\partial X_r})</td>
<td>(\frac{\partial Y}{\partial X_r} = AJB)</td>
<td>(\frac{\partial Y}{\partial X_e} = CJ^T D)</td>
</tr>
<tr>
<td>(\frac{\partial Y}{\partial X_e})</td>
<td>(\frac{\partial Y}{\partial X_e} = CJ^T D)</td>
<td>(\frac{\partial Y}{\partial X_e} = AJB)</td>
</tr>
</tbody>
</table>

\(^5\) If \(I\) is the identity matrix whose order is the number of rows of \(M\), we have \(M^{(k)} = I^{(k)} \cdot M\).
\[
\frac{\partial Y}{\partial \langle X \rangle} = \Sigma g A_g JB_g + \Sigma h C_h I^T D_h
\]  

(3.1)

we can state
\[
\frac{\partial Y_r}{\partial X_r} = \Sigma g A^T_g \otimes B_g + \Sigma h (C^T_h \otimes D_h)_{(m)}
\]  

(3.2)

\[
\frac{\partial Y_e}{\partial X_e} = \Sigma g B_g \otimes A^T_g + \Sigma h (D_h \otimes C^T_h)_{(n)}
\]  

(3.3)

e tc. (Compare [14, p. 956]). In the sequel we always use $\partial Y_r/\partial X_r$.

We illustrate the use of (3.2) in Table 3 for various forms of $Y$. Certain details are available in [6] and [7].

The vector derivative formulae of Table 1 may be obtained as special cases of the ones above. For example, if $x$ is a vector in (3.4),
\[
\frac{\partial y}{\partial \langle x \rangle} = A j.
\]

TABLE 3. FORMULAE FOR SOME COMMONLY OCCURRING FORMS OF $Y$ AND $\partial Y/\partial \langle X \rangle$

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$\frac{\partial Y}{\partial \langle X \rangle}$</th>
<th>$\frac{\partial Y_r}{\partial X_r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AX$</td>
<td>$AJ$</td>
<td>$A^T \otimes I_n$</td>
</tr>
<tr>
<td>$XA$</td>
<td>$JA$</td>
<td>$I_n \otimes A$</td>
</tr>
<tr>
<td>$AX^T$</td>
<td>$AJ^T$</td>
<td>$(A^T \otimes I_m)_{(m)}$</td>
</tr>
<tr>
<td>$X^T A$</td>
<td>$J^T A$</td>
<td>$(J^T A \otimes A)_{(m)}$</td>
</tr>
<tr>
<td>$AXB$</td>
<td>$AJ B$</td>
<td>$A^T \otimes B$</td>
</tr>
<tr>
<td>$AXB$</td>
<td>$A J^T B$</td>
<td>$(A^T \otimes B)_{(m)}$</td>
</tr>
<tr>
<td>$X^T X$</td>
<td>$J^T X + X^T J$</td>
<td>$(I_n \otimes X)_{(m)} + X \otimes I_n$</td>
</tr>
<tr>
<td>$XX^T$</td>
<td>$J X^T + X J^T$</td>
<td>$I_m \otimes X^T + (X^T \otimes I_m)_{(m)}$</td>
</tr>
<tr>
<td>$X^T AX$</td>
<td>$J^T AX + X^T A J^T$</td>
<td>$(I_n \otimes AX)_{(m)} + A^T X \otimes I_n$</td>
</tr>
<tr>
<td>$X AX^T$</td>
<td>$J A X^T + X A J^T$</td>
<td>$I_m \otimes AX^T + (A^T X^T \otimes I_m)_{(m)}$</td>
</tr>
<tr>
<td>$CX^T AXD$</td>
<td>$C(J^T AX + X^T A J^T)D$</td>
<td>$(C^T \otimes AXD)_{(m)} + A^T X^T C^T \otimes D$</td>
</tr>
<tr>
<td>$CXAX^T D$</td>
<td>$C(JAX^T + XAJ^T) D$</td>
<td>$C^T \otimes AX^T D + (A^T X^T C^T \otimes D)_{(m)}$</td>
</tr>
<tr>
<td>$X^{-1}$</td>
<td>$-X^{-1} VX^{-1}$</td>
<td>$-X^{-1} \otimes X^{-1}$</td>
</tr>
<tr>
<td>$AX^{-1}B$</td>
<td>$-AX^{-1} VX^{-1} B$</td>
<td>$-X^{-T} A^T \otimes X^{-1} B$</td>
</tr>
<tr>
<td>$(AXB)^{-1}$</td>
<td>$-(AXB)^{-1} A JB(AXB)^{-1}$</td>
<td>$-A^T (AXB)^{-T} \otimes B (AXB)^{-1}$</td>
</tr>
<tr>
<td>$C(AXB)^{-1} D$</td>
<td>$-C(AXB)^{-1} A JB(AXB)^{-1} D$</td>
<td>$-A^T (AXB)^{-T} C^T \otimes B (AXB)^{-1} D$</td>
</tr>
</tbody>
</table>

$A, B, C, D$ are not functions of $X$ in column 1. This restriction is not necessary in obtaining column 3 from column 2.
If \( x \) is a column vector, \( n \) is 1 and

\[
\frac{\partial y}{\partial x} = A^T \otimes 1,
\]

which is \( A^T \) as in (2.1). If \( x \) is a row vector, \( A^T \otimes I_n \) in (3.4) does not reduce any further. An interesting feature is observed when \( x \) in (3.7) is a column vector. Here \( n \) is 1, so

\[
\frac{\partial y}{\partial x} = (1 \otimes A)_{(m)} = A_{(m)}.
\]

But \( A \) has \( m \) rows, so \( A_{(m)} = A \) and we obtain (2.2). This feature occurs if we obtain (2.3), (2.4), (2.5) as respective special cases of (3.10), (3.12), (3.7) with \( x \) a column vector. When \( x \) is a row vector, a slightly different effect is sometimes noticed. As an example let \( x \) be a row vector in (3.6), so \( m \) is 1. Then

\[
\frac{\partial y}{\partial x} = (A^T \otimes 1)_{(1)} = A^T
\]

as in (2.1'), since \( M_{(0)} = M \) for any matrix \( M \).

4. SECOND ORDER DERIVATIVES USEFUL IN TESTING EXTREMA

There is no discussion in [6] regarding the testing of extrema using second order derivatives. Though sometimes this testing can be done using alternate approaches [4], [5], [15, pp. 448-449], [18], the matrix derivative theory should be extensive enough to provide those tests for extrema which feature second order matrix derivatives. A chief aim of this paper is the presentation of such theory.

A first important remark is the fact that the second order derivatives needed are first order derivatives of a derivative function which usually does not have any \( J \) or \( K \) in its value. Essentially the function to be maximized or minimized is a scalar function of the matrix elements, \( f = f(X) \). Usually the formal statement of the problem embodies this. Then \( \partial f/\partial X \) is a matrix expression giving all the \( mn \) derivatives \( \partial f/\partial \langle X \rangle_{as} \) without any \( J \) or \( K \). For example if

\[
\frac{\partial Y}{\partial \langle X \rangle} = \Sigma_{\alpha} A_\alpha J B_\alpha + \Sigma_{\alpha} C_\alpha J^T D_\alpha
\]

(4.1)

where \( A_\alpha, B_\alpha, C_\alpha \) and \( D_\alpha \) may be functions of \( X \), the methods of [6] show that, with \( Y = f \), scalar,

\[
\frac{\partial f}{\partial X} = \Sigma_{\alpha} A_\alpha^T B_\alpha + \Sigma_{\alpha} C_\alpha D_\alpha.
\]

(4.2)

The second order derivative terms we need are simply the derivatives of these \( mn \) terms with respect to the individual \( \langle X \rangle_{as} \). So we form the second order matrix derivative

\[
\frac{\partial}{\partial \langle X \rangle} \frac{\partial f}{\partial X}.
\]
which is a first order partial derivative of \( \partial f / \partial X \) featuring \( J \) and \( J^T \), and then present the results for

\[
\frac{\partial}{\partial X_r} \left[ \frac{\partial f}{\partial X} \right]^\epsilon
\]

in matrix form according to the rules of Section 3. We then apply the rules on continuous second order partial derivatives in testing for extrema: assuming differentiability of \( f \) at \( X_* \), a necessary condition for a relative minimum/maximum at \( X_* \) is, in formal matrix notation,

\[
\frac{\partial f}{\partial X} = 0 \quad \text{at} \quad X_*,
\]

and a sufficient condition is that in addition the Hessian matrix (of second order partial derivatives)

\[
\frac{\partial}{\partial X_r} \left[ \frac{\partial f}{\partial X} \right] = H
\]

be positive/negative definite at \( X_* \) [8, p. 62], [9, pp. 88–89]. The Hessian matrix is symmetric at \( X \) if the second order partial derivatives are continuous there [2, p. 152]. Some commonly occurring forms of Hessian matrices are discussed in Appendix B.

The properties above hold for the \( p^2 \times p^2 \) Hessian matrix resulting from differentiation of a square matrix with respect to \( \text{matrix} \) elements no matter whether or not equality relations (such as those of symmetry) exist among the scalar values of the elements, though any such relations may be applied to the general result after the differentiation is complete. The formulae throughout this paper are obtained by differentiation with respect to these matrix elements and provide general results which differ from the scalar element results when equality relations are present.

It is common practice to differentiate with respect to scalar elements but, when this is done, the results are different for every different set of equality relations. The equality relations encountered most frequently in statistics are those of symmetry with \( x_{\beta \alpha} = x_{\alpha \beta} \). If differentiation is with respect to scalar elements, the Hessian results (matrix elements) above must be modified, since

\[
\frac{\partial ( \ )}{\partial x_{\beta \alpha}} = \frac{\partial ( \ )}{\partial x_{\alpha \beta}}
\]

when \( x_{\beta \alpha} = x_{\alpha \beta} \). Hence the different rows identified by \( \alpha, \beta \) and \( \beta, \alpha \), with \( \beta \neq \alpha \), in the \( p^2 \times p^2 \) Hessian matrix (scalar elements) are identical, so this Hessian matrix is singular and not positive/negative definite at \( X_* \). Also the value of \( \partial f / \partial X \), is modified, since, while

\[
\frac{\partial X}{\partial (X)_\gamma} = J_{\gamma \delta}, \quad \frac{\partial X}{\partial (X)_{\delta \gamma}} = J_{\delta \gamma},
\]

* For \( f \) scalar, \( \partial f / \partial X_r = \partial f / \partial X_r \).
we also have
\[
\frac{\partial X}{\partial x_{\gamma \delta}} = \frac{\partial X}{\partial x_{\gamma \gamma}} = J_{\gamma \delta} + J_{\gamma \gamma}.
\]
It is then appropriate, if no additional equality relations exist in addition to those of symmetry, to use the
\[
\frac{p(p + 1)}{2} \times \frac{p(p + 1)}{2}
\]
Hessian matrix \(H'\) (scalar elements) which results from selecting columns \(\gamma, \delta\) with \(\delta \geq \gamma\) and rows \(\alpha, \beta\) with \(\beta \geq \alpha\). Then the elements of the \(\gamma, \delta\) column of \(H'\) are the sums of the corresponding \(\gamma, \delta\) and \(\delta, \gamma\) elements of \(H\).

With some effort it can be shown in particular cases that \(H'\) is positive/negative definite with \(H\). For example, if \(H\) is positive definite with form \(A \otimes A\) where \(p \times p\) symmetric, \(A\) is positive definite, then \(H''\), the
\[
\frac{p(p + 1)}{2} \times \frac{p(p + 1)}{2}
\]
matrix which results from combining each \(\beta, \alpha\) row with the \(\alpha, \beta\) row, each \(\delta, \gamma\) column with the \(\gamma, \delta\) column, is positive definite since its quadratic form is a special case of the quadratic form of \(H\). Comparing with the rows of \(H'\) it is seen that the \(\alpha, \alpha\) row of \(H'\) is equal to the corresponding row of \(H''\) and that the \(\alpha, \beta\) row of \(H'\) equals \(\frac{1}{2}\) that of \(H''\). Then \(H'\) is positive definite since its principal minors are positive with the corresponding principal minors of \(H''\).

However, such effort is unnecessary since results are immediately available from \(H\), which has no singularities because of equal rows resulting from differentiations with respect to equal elements. Matrix elements, rather than (conventional) scalar elements, are used throughout this paper to obtain formulae which simultaneously attain generality, simplicity, and applicability.

Of course, if it is known that the function is convex (or concave) then it is not necessary to apply the Hessian conditions. One can use the theorem [8, p. 62] that if \(f\) is differentiable and convex/concave on an open convex set \(A\) and \(\partial f/\partial X = 0\) at \(X_0 \in A\), then \(f\) has an absolute minimum/maximum at \(X_0\). A referee points out that a function \(f(X)\) is convex if and only if \(f(\lambda X + (1 - \lambda) Y)\) is convex in \(\lambda\), so that in appropriate problems, differentiation can be performed with respect to the scalar \(\lambda\).

For a general treatment of the extrema problem with constraints, we follow Gillespie [9] and use vector differentiation. Let \(f = f(x) = f(x_1, \cdots, x_n)\) be the (scalar) function to be extremized subject to the \(m < n\) constraints \(g_i(x) = 0, \quad i = 1, \cdots, m\). With \(\lambda_i\) as the Lagrange multiplier of the \(i\)th constraint, the function to be extremized is
\[
\phi(x) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)
\]
which, on using vectors \(\lambda, g\) as \(m\)-component column vectors, may be expressed as
\[
\phi(x) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)
\]
\[ \phi = f + \lambda^T g. \] (4.4)

The theory above is used by differentiating (4.4) with respect to the unknown vectors \( x, \lambda \) and equating to the zero vector. The necessary conditions at the solution \( s \) are

\[ \frac{\partial \phi}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x} \lambda = 0 \] (4.5)

\[ \frac{\partial \phi}{\partial \lambda} = g = 0. \] (4.6)

Let \( x_s, \lambda_s \) denote the values of \( x, \lambda \) which satisfy the above equations. Differentiating again we get the Hessian matrix

\[ H = \begin{bmatrix}
\frac{\partial}{\partial x} \frac{\partial \phi}{\partial x} & \frac{\partial}{\partial x} \frac{\partial \phi}{\partial \lambda} \\
\frac{\partial}{\partial \lambda} \frac{\partial \phi}{\partial x} & \frac{\partial}{\partial \lambda} \frac{\partial \phi}{\partial \lambda}
\end{bmatrix} = \begin{bmatrix}
\frac{\partial}{\partial x} \frac{\partial \phi}{\partial x} & \frac{\partial}{\partial x} g \\
\frac{\partial}{\partial x} g^T & 0
\end{bmatrix}. \] (4.7)

Let the Hessian matrix at \( x_s \) be denoted by \( H_s \). Then a sufficient condition for \( f \) to have a constrained minimum/maximum at \( x_s \) is that

\[ \frac{\partial}{\partial x} \frac{\partial \phi}{\partial x} \]

subject to the linear conditions

\[ (x - x_s)^T \frac{\partial g}{\partial x} = 0, \]

be positive/negative definite.\(^7\) Thus, we investigate the nature of

\[ \frac{\partial}{\partial x} \frac{\partial \phi}{\partial x}, \]

on a subspace of \( E^n \). Some working rules, given in [9, pp. 97–98], which depend on the signs of certain principal minors of \( H_s \), are helpful in applying this principle to problems. These are presented in Appendix C.

5. APPLICATIONS

We illustrate the theory by applying it to several different types of problems. These applications deal with (a) extrema of quadratic forms, (b) general least squares theory, and (c) maximum likelihood estimation of the parameters of the multivariate normal and such generalizations as multivariate linear regression, including the case of subgroups with missing observations [17].

\(^7\) The reader may wish to compare this with Fleming [8, p. 134, ex. 11]. Goldberger [10, p. 47] applies the second order condition on \( f \) (subject to the side conditions) instead of \( \phi \). If the constraints \( g \) are linear in \( x \) as is common in applications, then

\[ \frac{\partial}{\partial x} \frac{\partial g}{\partial x} = 0 \quad \text{and consequently} \quad \frac{\partial}{\partial x} \frac{\partial \phi}{\partial x} = \frac{\partial}{\partial x} \frac{\partial f}{\partial x}. \]
(a) Extrema of Quadratic Forms

Bush and Olkin [4, p. 485], [5, p. 474] have shown that many applied problems in multivariate analysis are special cases of the maximization of the ratio of two quadratic forms or the minimization of a quadratic form subject to linear restrictions. An essential element in their proofs is the use of the Cauchy-Schwarz inequality. It is our claim that matrix (vector) differentiation can also be used to solve problems of this type directly as illustrated below.

(i) We wish to find the absolute maximum (supremum) of the ratio $x^T B x / x^T A x$ where $A$ and $B$ are symmetric and $A$ is positive definite. Letting $f$ be this ratio, we have $f \cdot x^T A x = x^T B x$. Differentiating with respect to $x$ we get

$$2f A x + \frac{\partial f}{\partial x} \cdot x^T A x = 2B x$$

so that

$$\frac{\partial f}{\partial x} = \frac{2(B - fA)x}{x^T A x}.$$  

Since $x^T A x > 0$ for all $x \neq 0$, we get, as the necessary condition for any extremum, $(B - fA)x = 0$. Hence, it is evident immediately that $f$ is the characteristic root of $BA^{-1}$, and we have for $BA^{-1}$ positive definite,

$$\lambda_m \leq \frac{x^T B x}{x^T A x} \leq \lambda_M$$

where $\lambda_m$ is the smallest characteristic root and $\lambda_M$ the largest. No second order derivatives are necessary in this case because of the well known properties of the characteristic roots. The characteristic vector $x_m / \lambda_m$ associated with $\lambda_M / \lambda_m$ gives the supremum/infimum of $f$ in all cases in which the characteristic roots are not all the same.

There is no local maximum/minimum associated with a multiple $\lambda$. For example, with $B = I$ and $A = I$, $f = 1$, all $\lambda_i = 1$ and there is no maximum, minimum, infimum nor supremum. However, if there are at least two different values of $\lambda_i$ then $\lambda_m < \lambda_M$ so there is an infimum and a supremum, but no local maximum/minimum if all $\lambda_i$ are multiple.

Rao [15, p. 50] gives the special result when $A = I$. Bush and Olkin [4, p. 485] and Rao [15, p. 48] give the special results when $B$ is of rank 1 with $B = uu^T$. Then there is only one positive root, $\lambda_M$, of $|B - \lambda A| = 0$. Hence

$$f = \frac{x^T B x}{x^T A x} = \frac{(u^T x)^2}{x^T A x} \leq \lambda_M.$$  

Now $|uu^T - \lambda A| = (-\lambda)^{p-1} |A| (u^T A^{-1} u - \lambda)$ and $\lambda_M = u^T A^{-1} u$, or see [1, p. 108]. Then $x_* = A^{-1} u$ is the solution since $x_*^T A x_* = u^T \lambda = u^T A^{-1} u$ and $\lambda_M$ is attained.

(ii) We seek the absolute minimum/maximum of $x^T A x$ subject to the consistent restrictions $B^T x = u$ where the positive definite $A$ is $p \times p$, $B$ is $p \times k$ with $k < p$, $x$ is $p \times 1$ so $u$ is $k \times 1$. Solutions have been provided by Bush and Olkin [4, p. 485], [5, p. 474] and Rao [15, p. 49] who use the Cauchy-Schwarz in-
equality. We follow Rao in requiring consistency, but not necessarily full column rank of $B$. Our calculus proof, using Lagrange multipliers for the side conditions, is more like that of Goldberger [10, pp. 47–48].

With $2\lambda^T$ as the Lagrange multiplier vector, where $\lambda$ is $k \times 1$, we seek the minimum of

$$
\phi = x^T A x + 2\lambda^T (u - B^T x).
$$

(5.3)

Differentiating with respect to $x$, $\lambda$, we get

$$
\frac{\partial \phi}{\partial x} = 2Ax - 2B\lambda, \quad \frac{\partial \phi}{\partial \lambda} = u - B^T x
$$

and the necessary condition for a solution $x_*$ is $Ax_* = B\lambda_*$, where $u = B^T x_*$. Then $x_* = A^{-1}B\lambda_*$, so $B^T x_* = B^T A^{-1}B\lambda_*$ and

$$
\lambda_* = (B^T A^{-1}B)^{-1}B^T x_* = (B^T A^{-1}B)^{-1}u
$$

(5.4)

where $(B^T A^{-1}B)^{-1}$ is the generalized inverse [15, p. 24] and becomes $(B^T A^{-1}B)^{-1}$ when $B$ is of full column rank. The solution is

$$
x_* = A^{-1}B(B^T A^{-1}B)^{-1}u.
$$

(5.5)

In the second order condition we note that

$$
\frac{\partial}{\partial x} \frac{\partial \phi}{\partial x} = 2A^T = 2A
$$

which is positive definite, i.e. the quadratic form $(x - x_*)^T (2A) (x - x_*) > 0$ for all $x \neq x_*$. This set includes $x$ such that

$$
(x - x_*)^T \frac{\partial \phi}{\partial x} = 0
$$

or $(x - x_*)^T B = 0$.

Here $x_*$ provides a minimum if $A$ is positive definite and it would provide a maximum if $A$ were negative definite. It is interesting to note that the nature of the solution is entirely determined by $A$, and not at all by the matrix $B$ of linear constraints.

It is not the purpose here to argue that these calculus methods are better for all problems or that they should necessarily replace methods featuring inequalities, gradients, or convexity properties in common use. Our argument is simply that the direct fundamental calculus approach is commonly available for generalizations of these problems with appropriate simple differentiation formulae. It is recognized that generalizations of the type indicated above are not necessarily of interest in statistics which features matrices, such as covariance matrices, which are Gramian. However, we reemphasize that easy

* Here

$$
\frac{\partial \phi}{\partial x} = -B, \quad \text{though} \quad \frac{\partial}{\partial x} \frac{\partial \phi}{\partial \lambda} = -2B.
$$

since we use $2\lambda^T$ as Language multiplier (cf. (4.7)).
calculus methods are available for those who wish to use them. Rao, in the summary of results for extrema of quadratic forms [15, pp. 48–53], makes no mention of the possible use of the calculus, either in the text or in the exercises, in obtaining results such as those above, though he does give in an exercise [15, p. 57] formulae for

\[ \frac{\partial}{\partial x} (B^T x), \quad \frac{\partial}{\partial x} (x^T x), \quad \frac{\partial}{\partial x} (x^T A x), \]

with an application to least squares.

In the illustrations following we take up the general problems involving matrices as well as vectors.

(b) General Least Squares Theory

We denote the error by \( e = y - Xb \), so that the least squares criterion calls for the minimization of \( w = e^T e \). Here

\[ \frac{\partial w}{\partial b} = -2X^T y + 2X^T X b, \]

so the solution is \( b = (X^T X)^{-1} X^T y \) for \( X^T X \) positive definite. Using (2.1),

\[ \frac{\partial}{\partial b} \frac{\partial w}{\partial b} = 2X^T X \]

which is positive definite, establishing that \( b \) minimizes \( w \).


\[ \frac{\partial \ln |W|}{\partial B} = \sum_{\gamma, \delta} \langle W^{\gamma,\delta} \rangle \frac{\partial \langle W \rangle_{\gamma,\delta}}{\partial B} \]

\[ = \sum_{\gamma, \delta} \langle W^{\gamma,\delta} \rangle (-X^T Y + X^T XB)(K_{\gamma,\delta} + K_{\gamma,\delta}^T) \]

\[ = 2(X^T XB - X^T Y)W^{-1} \quad (5.6) \]

since \( W \) is symmetric. With \( X^T X \) positive definite, the solution is \( \hat{b} = (X^T X)^{-1} X^T y \), which is a generalization of \( b \) of the previous problem. The \( \hat{b}_i = (X^T X)^{-1} X^T y_i \) which minimize the diagonal elements \( (E^T E)_{ii} = e_i^T e_i \) of \( W \) collect to form \( \hat{B} \) which minimizes \( \ln |W| \) [10, p. 205]. We obtain on differentiating (5.6),

\[ \frac{\partial}{\partial \langle B \rangle} \frac{\partial \ln |W|}{\partial B} = 2X^T X W^{-1}, \]

so that

\[ \frac{\partial}{\partial B_r} \frac{\partial \ln |W|}{\partial B_r} = 2X^T X \otimes W^{-1}, \]
which is positive definite, thus establishing that $\hat{B}$ minimizes $\ln|W|$ as well as the diagonal terms of $W$.

(c) Multivariate Normal

(i) We consider the maximum likelihood estimates of $\mu$ and $\Sigma$ for the $p$-variate normal and wish to establish formally that they indeed provide a maximum for the likelihood function. Differentiation of

$$\ln L = \text{const} - \frac{N}{2} \ln |\Sigma| - \frac{1}{2} \Sigma(x - \mu)^T \Sigma^{-1}(x - \mu)$$

with respect to the matrix elements of $\mu$ and $\Sigma$ gives, (see [6, p. 618] for details),

$$\frac{\partial \ln L}{\partial \mu} = \Sigma^{-1} \Sigma(x - \mu)$$

(where one can use

$$\frac{\partial \ln L}{\partial \mu} = \frac{\partial \ln L}{\partial (x - \mu)} \frac{\partial (x - \mu)}{\partial \mu}$$

or expand the last term of (5.7))

$$\frac{\partial \ln L}{\partial \Sigma} = -\frac{N}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} \{\Sigma(x - \mu)(x - \mu)^T\} \Sigma^{-1}$$

$$= -\frac{N}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} B \Sigma^{-1}, \text{ say}$$

with solution $s$ of the likelihood equations

$$\hat{\mu} = \bar{x}, \quad \hat{\Sigma} = \frac{1}{N} \Sigma(x - \bar{x})(x - \bar{x})^T.$$  

(5.10)

Throughout 5(c) we assume that $\Sigma$ as well as its estimate $\Sigma$ is positive definite. Differentiating again, using (2.1), (3.16), (3.17),

$$\left[ \frac{\partial}{\partial \mu} \frac{\partial \ln L}{\partial \mu} \right]_s = -N \Sigma^{-1} - \frac{N}{2} \Sigma^{-1} B \Sigma^{-1}$$

and

$$\left[ \frac{\partial}{\partial \Sigma} \frac{\partial \ln L}{\partial \Sigma} \right]_s = -\frac{N}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} B \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} B \left[ \frac{\partial \Sigma^{-1}}{\partial \Sigma} \right]_s$$

$$= \frac{N}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} B \Sigma^{-1} - \frac{1}{2} \Sigma^{-1} B \Sigma^{-1}$$

$$= -\frac{N}{2} \Sigma^{-1}$$

since $B]_s = N \hat{\Sigma}$ and $\Sigma^{-1}]_s = \hat{\Sigma}^{-1}$. Thus
\[
\frac{\partial}{\partial \Sigma_r} \frac{\partial \ln L}{\partial \Sigma_r} = - \frac{N}{2} \hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{-1}.
\]

At the solution \( s \), the mixed partials are 0, so the Hessian matrix is
\[
\begin{pmatrix}
-\frac{N}{2} \hat{\Sigma}^{-1} & 0 \\
0 & -\frac{N}{2} \hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{-1}
\end{pmatrix},
\]
which is negative definite (B.2), establishing that \( \hat{\mu}, \hat{\Sigma} \) maximize \( \ln L \). When \( \mu \) or \( \Sigma \) is known, the Hessian matrix at the solution is
\[-\frac{N}{2} \hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{-1}\]
or \(-N\Sigma^{-1}\) respectively, each of which is negative definite, thus establishing that
\[
\hat{\Sigma} = \frac{1}{N} \Sigma_a (x_a - \mu)(x_a - \mu)^T,
\]
or \( \hat{x} \), is indeed the maximum likelihood estimate of the unknown parameter.

An interesting alternative to the use of matrix derivatives in establishing this property was given by Watson [18] and is outlined by Rao [15, pp. 448–449]. Some feel that concavity of \( \ln L \) in \( \mu \) and \( \Sigma^{-1} \) is so well known that no further treatment is necessary. However, this concavity is easily established if we differentiate (5.7) with respect to \( \mu \) and \( \Sigma^{-1} \) (see [1, p. 48] for justification) and use the results of [6, p. 617] with \( \ln |\Sigma| = -\ln |\Sigma^{-1}| \) and
\[
\frac{\partial}{\partial \Sigma^{-1}} \ln |\Sigma^{-1}| = \Sigma.
\]

Then we get (5.8) and, in place of (5.9), immediately
\[
\frac{\partial \ln L}{\partial \Sigma^{-1}} = \frac{N}{2} \Sigma - \frac{1}{2} \Sigma_a (x_a - \mu)(x_a - \mu)^T
\]
(5.9')
yielding the same solution as (5.10). So (5.11) is still applicable and using (3.16),
\[
\frac{\partial}{\partial (\Sigma^{-1})} \bigg[ \frac{\partial \ln L}{\partial \Sigma^{-1}} \bigg] = - \frac{N}{2} \Sigma J \Sigma \bigg] = - \frac{N}{2} \hat{\Sigma} J \hat{\Sigma}
\]
(5.12')
so that
\[
\frac{\partial}{\partial \Sigma_r} \frac{\partial \ln L}{\partial \Sigma_r} = - \frac{N}{2} \hat{\Sigma} \otimes \hat{\Sigma}.
\]
The mixed partials at \( s \) are 0, so the Hessian there with respect to \( \mu \) and \( \Sigma^{-1} \) is negative definite (B.2), establishing concavity of \( \ln L \).

It may be noted here that the last term of (5.7) is linear in \( \Sigma^{-1} \) and hence
makes no contribution to (5.12'). This is similar to the situation of the quadratic form problem in which the linear restrictions have no influence on the second derivative.

(ii) We next consider maximum likelihood estimation of parameters \( \beta, \Sigma \) in multivariate linear regression when \( N \) observations \( x_a \) are drawn from a \( N(\beta z_a, \Sigma) \) population [1, p. 179]. Here \( \ln L \) is obtained by replacing \( \beta z_a \) for \( \mu \) in (5.7), and using the chain rule

\[
\frac{\partial \ln L}{\partial \beta} = \frac{\partial \ln L}{\partial (x_a - \beta z_a)} \frac{\partial (x_a - \beta z_a)}{\partial \beta}
\]

or expanding \((x_a - \beta z_a)^T \Sigma^{-1} (x_a - \beta z_a)\), we get

\[
\frac{\partial \ln L}{\partial \beta} = \Sigma^{-1} \Sigma_a (x_a - \beta z_a) z_a^T,
\]

while \( \partial \ln L / \partial \Sigma \) is the same as (5.9) with \( \beta z_a \) for \( \mu \). The solution is then

\[
\hat{\beta} = (\Sigma_a x_a z_a^T (\Sigma_a x_a z_a^T)^{-1})^{-1},
\]

with \( \Sigma_a z_a^T \) non-singular, and

\[
\hat{\Sigma} = \frac{1}{N} \Sigma_a (x_a - \hat{\beta} z_a) (x_a - \hat{\beta} z_a)^T.
\]

Differentiating again,

\[
\frac{\partial}{\partial (\beta)} \left[ \frac{\partial \ln L}{\partial \beta} \right]_{s} = -\hat{\Sigma}^{-1} J \Sigma_a z_a^T,
\]

while

\[
\frac{\partial}{\partial \Sigma} \left[ \frac{\partial \ln L}{\partial \Sigma} \right]_{s}
\]

is the same as (5.12) (\( \beta z_a \) replaces \( \mu \)). The mixed derivatives are 0 at \( s \), so the Hessian matrix there is

\[
\begin{pmatrix}
-\Sigma^{-1} \otimes \Sigma_a z_a z_a^T & 0 \\
0 & -\frac{N}{2} \hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{-1}
\end{pmatrix},
\]

which is negative definite when \( \hat{\Sigma}^{-1} \) and \( \Sigma_a z_a z_a^T \) are positive definite, establishing concavity. (As in (i), differentiation can be performed more simply with respect to \( \beta \) and \( \Sigma^{-1} \).)

(iii) Trawinski and Bargmann [17] consider the general linear hypothesis featuring \( K \) groups of \( n \) units each, where in each group observations are available for a different subset of size \( u \) out of the \( p \) variables under study. \( Y_i^T \) denotes the \( n \times u \) matrix of observations in the \( i \)th group, \( i = 1, \ldots, K \). If \( Y_i^T \) denotes the \( n \times p \) matrix of theoretically possible observations, and \( M_i \) is a \( p \times u \) matrix of appropriately placed ones and zeros, \( Y_i^T = Y_i^T M_i \). The model considered is \( E(Y_i^T) = A \xi M_i \), where \( A \) is the common \( n \times m \) design matrix and \( \xi \) the \( m \times p \) parameter matrix. If \( \Sigma \) is the common \( p \times p \) covariance matrix for the \( p \) variables and \( U_i \) the \( u \times u \) covariance matrix for each row vector of \( Y_i^T \),
MULTIVARIATE MAXIMA AND MINIMA WITH MATRIX DERIVATIVES

\( U_i = M_i^T \Sigma M_i \). If we let \( P_i = Y_i - M_i^T \xi \xi^T A^T \), the log likelihood is

\[
\ln L = \text{const} - \frac{n}{2} \sum_{i=1}^K \ln |U_i| - \frac{1}{2} \text{tr} \sum_{i=1}^K U_i^{-1} P_i P_i^T. \tag{5.13}
\]

It is then shown in [17], and with the use of matrix derivatives in [6, p. 619], that

\[
\frac{\partial \ln L}{\partial \xi} = \sum_{i=1}^K A^T P_i^T U_i^{-1} M_i^T = \sum_{i=1}^K A^T (Y_i^T - A \xi M_i) U_i^{-1} M_i^T \tag{5.14}
\]

\[
\frac{\partial \ln L}{\partial \Sigma} = - \frac{n}{2} \sum_{i=1}^K M_i U_i^{-1} M_i^T + \frac{1}{2} \sum_{i=1}^K M_i U_i^{-1} P_i P_i^T U_i^{-1} M_i^T. \tag{5.15}
\]

The necessary conditions are obtained by setting (5.14) and (5.15) equal to 0. Then at a solution, using (3.8)

\[
\frac{\partial}{\partial \langle \xi \rangle} \frac{\partial \ln L}{\partial \xi} = - A^T A J \sum_{i=1}^K M_i \mathcal{O}_i^{-1} M_i^T \tag{5.16}
\]

and with \( U_i = M_i^T \Sigma M_i \) in (5.15) together with the use of (3.18)

\[
\frac{\partial}{\partial \langle \Sigma \rangle} \frac{\partial \ln L}{\partial \Sigma} = - \frac{n}{2} \sum_{i=1}^K M_i U_i^{-1} M_i^T J M_i U_i^{-1} M_i^T
\]

\[
+ \frac{1}{2} \sum_{i=1}^K M_i U_i^{-1} M_i^T J M_i U_i^{-1} P_i P_i^T U_i^{-1} M_i^T
\]

\[
+ \frac{1}{2} \sum_{i=1}^K M_i U_i^{-1} P_i P_i^T U_i^{-1} M_i^T J M_i U_i^{-1} M_i^T \tag{5.17}
\]

\[
= - \frac{n}{2} \sum_{i=1}^K M_i \mathcal{O}_i^{-1} M_i^T J M_i \mathcal{O}_i^{-1} M_i^T,
\]

since \( P_i P_i^T \langle \xi \rangle = n \mathcal{O}_i \langle U_i^{-1} \rangle = \mathcal{O}_i^{-1} \). The mixed partial derivatives are 0, and the Hessian matrix is

\[
\begin{bmatrix}
- A^T A \otimes \sum_{i=1}^K M_i \mathcal{O}_i^{-1} M_i^T & 0_{mp \times p^2} \\
0_{p^2 \times mp} & - \frac{n}{2} \sum_{i=1}^K M_i \mathcal{O}_i^{-1} M_i^T \otimes M_i \mathcal{O}_i^{-1} M_i^T
\end{bmatrix} \tag{5.18}
\]

With \( \mathcal{O}_i^{-1} \) positive definite and \( M_i \) a \( p \times u \) matrix consisting of an \( I_u \) modified by the addition of \( p-u \) zero rows, \( M_i \mathcal{O}_i^{-1} M_i^T \) is positive semidefinite and so is

\[
\sum_{i=1}^K M_i \mathcal{O}_i^{-1} M_i^T.
\]

But

\[
\sum_{i=1}^K M_i \mathcal{O}_i^{-1} M_i^T = 0
\]

implies \( \mathcal{O}_i^{-1} = 0 \). Thus
\[
\sum_{i=1}^{K} M_i U_i^{-1} M_i^T
\]
is positive definite. Similar remarks hold true for
\[
\sum_{i=1}^{K} M_i U_i^{-1} M_i^T \otimes M_i U_i^{-1} M_i^T.
\]
So (5.18) is negative definite, establishing concavity at a solution, even though a formal explicit solution is not provided.

The use of \( \Sigma^{-1} \) in place of \( \Sigma \), which enables us to reduce the second order derivative in the earlier problem to a relatively simple form (5.12'), complicates the results in this case. The differentiation with respect to \( U_i^{-1} \) is relatively simple, but since the resulting likelihood equations are inconsistent, it does not lead to a common estimate of \( \Sigma \). Considering the nature of the problem and that no one, to our knowledge, has established the character of the extrema in the general case with other methods, we feel that the use of matrix derivatives is appropriate and effective.

ACKNOWLEDGMENTS

The authors are indebted to David Culver, E. H. Jebe, Charles Kowalski, H. Neudecker, R. P. Singh, and Martin Weinrich who made valuable suggestions for improvement. They also wish to thank the referees and editors who were most helpful in providing information and suggestions for the final revision.

REFERENCES

APPENDIX A

SOME PROPERTIES OF KRONECKER PRODUCTS

1. \((A \otimes B)^T = A^T \otimes B^T\). \hspace{1cm} (A.1)

2. \((A \otimes D)(C \otimes D) = (AC) \otimes (BD)\), if \(AC, BD\) exist. \hspace{1cm} (A.2)

   Corollary: \((A \otimes B)^{-1} = A^{-1} \otimes B^{-1}\), if inverses exist \hspace{1cm} (A.3)
   
   since \((A \otimes B)(A^{-1} \otimes B^{-1}) = I \otimes I = I\).

3. If \(A, B\) are positive definite, then \(A \otimes B\) is positive definite.

APPENDIX B

SOME COMMONLY OCCURRING FORMS OF HESSIAN Matrices

The Hessian matrix often takes at the critical value, the following forms, or their generalizations, with a plus or minus sign:

\[
\begin{align*}
A \otimes B, \\
\begin{pmatrix} A \otimes B & 0 \\ 0 & C \otimes D \\ \end{pmatrix}, \\
\begin{pmatrix} A \otimes B & E \otimes F \\ E^T \otimes F^T & C \otimes D \\ \end{pmatrix},
\end{align*}
\]  \hspace{1cm} (B.1) \hspace{1cm} (B.2) \hspace{1cm} (B.3)

where the matrices \(A, B, C, D\) are known to be symmetric and positive definite.

Clearly (B.1) is positive definite being the Kronecker product of positive definite matrices (Appendix A), \([3, \text{p. 235}]\).

With \(A \otimes B\) positive definite and \(C \otimes D\) positive definite, (B.2) is positive definite since its associated quadratic form is the sum of the respective quadratic forms associated with \(A \otimes B\) and \(C \otimes D\).

If one obtains the form (B.3), its determinant is evaluated \([1, \text{p. 344}]\) as

\[
| A \otimes B - (E \otimes F)(C \otimes D)^{-1}(E^T \otimes F^T) | | C \otimes D |
\]

\[
= | A \otimes B - (E \otimes F)(C^{-1} \otimes D^{-1})(E^T \otimes F^T) | | C \otimes D | \hspace{1cm} \text{by (A.3)}
\]

\[
= | A \otimes B - EC^{-1}E^T \otimes FD^{-1}F^T | | C \otimes D | . \hspace{1cm} \text{by (A.2)}
\]

Similar evaluation of the naturally ordered principal minors of (B.3) can be carried out to test its positive definiteness.

APPENDIX C

RULES FOR DETERMINING CONSTRAINED EXTREMA

It is indicated in Section 4 that the function \(f\) has a constrained minimum/maximum at \(x\), if
\[
\frac{\partial^2 \phi}{\partial x_j \partial x_k}, \quad \text{and} \quad \frac{\partial^2 \phi}{\partial x_j \partial x_k}, \quad \text{and} \quad g_{ij} \quad \text{for} \quad \frac{\partial g_i}{\partial x_j}.
\]

Then the second order matrix of \( \phi \) with respect to \( x, \lambda \), evaluated at \( x_s \), may be written [9, pp. 97–98] as

\[
H_s = H = \\
\begin{pmatrix}
\phi_{11} & \phi_{12} & \cdots & \phi_{1n} & g_{11} & \cdots & g_{m1} \\
\phi_{21} & \phi_{22} & \cdots & \phi_{2n} & g_{12} & \cdots & g_{m2} \\
& & & & & & \\
& & & & & & \\
\phi_{n1} & \phi_{n2} & \cdots & \phi_{nn} & g_{1n} & \cdots & g_{mn} \\
g_{11} & g_{12} & \cdots & g_{1n} & 0 & \cdots & 0 \\
& & & & & & \\
& & & & & & \\
g_{m1} & g_{m2} & \cdots & g_{mn} & 0 & \cdots & 0
\end{pmatrix}.
\]

This simple notation is chosen with a view to use rather than specific detail. It should be remembered in the discussion below that \( H \) is a matrix of second order derivatives each of which is evaluated at \( x = x_s \).

Modifying Gillespie [9, p. 97] we let \( \Delta_0 = |H| \), \( \Delta_1 = |H_{11}| \) where \( H_{11} \) is \( H \) with the first column and the first row deleted. More generally, \( \Delta_i \) is \( |H_{ii}| \) where \( H_{ii} \) is obtained by deleting the first \( i \) rows and columns of \( H \). The process continues up to \( i = n - m - 1 \). Sufficient second order conditions for the maximization or minimization at \( x_s \), subject to the vanishing of the linear term in the Taylor expansion,

\[
(x - x_s)^{\tau} \frac{\partial g}{\partial x_s},
\]

with \( 0 \leq i \leq n - m - 1 \), are

i. for minimization, \( \Delta_i \geq 0 \) has the sign of \((-1)^n\)

ii. for maximization, \( \Delta_i \leq 0 \) has the sign of \((-1)^{n-i}\).

(C.2)