Multivariate Behavioral Research, 1985, 20, 113–139

On the Formal Differentiation of Traces and Determinants

Peter H. Schönemann
Purdue University and Universität Bielefeld

A compact notation for obtaining and handling matrices of partial derivatives is suggested in an attempt to generalize "symbolic vector differentiation" to matrices of independent variables. The proposed technique differs from methods advocated by Dwyer and MacPhail (1948) and Wrobleski (1963) in several respects, notably in a deliberate limitation on the classes of scalar functions considered: traces and determinants. To narrow interest to these two classes of scalar matrix functions allows one to invoke certain algebraic identities which simplifies the problem, because (a) the treatment of traces of products of matrices can be reduced to that of a few representatives of large equivalence classes of such products, all having the same formal derivative, and because (b) the more involved task of differentiating determinants of matrix products can be translated into the more amenable problem of differentiating the traces of such products. A number of illustrative examples are included in an attempt to show that the above limitation is not as serious as might at first appear, because traces and determinants apply to a wide range of psychometric and statistical problems.

In this paper an attempt will be made to motivate a possible compact notation for what shall be called "formal matrix differentiation"[1] of certain scalar matrix functions. By this is meant the partial differentiation of such functions with respect to each of the elements of a matrix and the subsequent rearrangement of these partial derivatives in a matrix of the same order.

The need for an efficient notation has been recognized before, especially with vectors as independent variables, although the treatment it has been accorded in the literature is sporadic and often sketchy (Aitken, 1959; Anderson, 1958; Bargmann, 1957; Bargmann & Mah, 1961; Bock & Bargmann, 1966; Cooley & Lohnes, 1962; Graybill,

[1]The term "formal", rather than "symbolic" was chosen in sympathy with a remark by Wrobleski (1963, p. 5): "It is this author's feeling that the adjective "symbolic" imputes a secondary or colored meaning to this basic mathematical notion of a matrix derivative of a scalar function of matrix variables . . ." After having made his choice of terminology the present writer was somewhat disheartened to read Bellman's semantic appraisal: "By the term "formal" we mean with attention to the rigorous aspects such as the existence of partial derivatives, interchange of limits, and so on. It is a fine word to cover a multitude of mathematical sins." (1961, p. 58 ff.).

Editor's Note: The contents of this paper first appeared in a University of North Carolina Research Bulletin in 1965. It contains important basic results on matrix derivatives, and pays special attention to their application to problems in psychometrics. Many papers have utilized these elegant results, and, although much subsequent work on matrix derivatives has appeared (see Nel, 1980, for an excellent summary), this paper remains, in an important sense, undated. We are delighted that Professor Schönemann agreed to add it to the literature. Professor Schönemann is currently in Germany, on sabbatical leave from Purdue University. His work abroad is receiving support from the Deutsche Forschungsgemeinschaft.

1961; Kaiser & Dickman, 1959; McKeon, 1962; Rao, 1952; Schönemann, 1965, 1966). It is anticipated, however, that interest in such efforts is likely to increase in the social sciences in proportion with the general increase of interest in elementary matrix methods, as it continues to be promoted through modern computing equipment.

One of the better known, and perhaps most systematic, efforts in this field is embodied in a paper by Dwyer and MacPhail (1948). This work has been elaborated upon since, notably in a report by Wrobleski (1963). Wrobleski attempts to put the original work of Dwyer and MacPhail on what the present author can only surmise to be a more respectable mathematical basis (invoking such concepts as Frechet-Gateaux differentials, Banach algebra, Hausfdorf differentials). Moreover, he generalizes this work in some of its more practical aspects, notably by extending the technique to include higher order derivatives.

The present efforts are more modest than Wrobleski's, and indeed, more modest than Dwyer and MacPhail's. In a rough sense the present paper aims in the opposite direction (relative to Wrobleski's) in that it specializes, rather than generalizes Dwyer and MacPhail's original work. By trading some generality it is hoped to gain some practicality.

To be more specific, the derivatives in this paper will always be in matrix form, without further rearrangements as are sometimes necessary in Dwyer and MacPhail's technique. Nor will there be any need for dummy matrices, (of type "J" or "K") which figure quite prominently in Dwyer and MacPhail's, and also in Wrobleski's work. Sometimes the final removal of such dummy matrices can be nearly as bothersome as the original problem might have been, had it been left in summation notation. To illustrate, consider the case where the scalar function of the (independent) matrix is a determinant (possibly of a matrix product). In this case, which as will be seen is of no small practical interest, Dwyer and MacPhail's technique provides convenient means to express the formal derivative of each element of the matrix of the determinant with respect to the independent matrix. Each one of these derivatives will be a matrix, possibly involving one or more dummy matrices.

These individual derivatives, one might hope, somehow relate to the derivative of the original function of the elements, i.e., to that of the determinant. But the problem of putting Humpty Dumpty together again is left to the ingenuity of the user. Possibly a more polite way of saying this would be to note that the Dwyer and MacPhail technique is impractical for determinants. Considering that every maximum likelihood problem based upon the multivariate normal distribution (or

Wishart distribution, for that matter) involves determinants, one comes to be less impressed with the apparent generality of the Dwyer and MacPhail technique.

On the other hand, the presently proposed technique has its obvious limitation in terms of the scalar functions considered. But the contention is that those functions chosen, traces and determinants, cover a wide range of applications, statistical as well as psychometric. As is well known, these two classes of scalar functions are not unrelated but rather constitute but two members of a larger class of (symmetric) functions and as such share certain properties which will be found convenient for the present purpose. Traces, in particular, are very simple functions of the (diagonal) elements of a square matrix. Notwithstanding their simplicity, traces provide a powerful tool for the formulation of least squares problems. (e.g. Bargmann & Mah, 1961; Cattell, 1944; Cliff, 1966; Eckart & Young, 1936; Edgerton & Kolbe, 1936; Gibson, 1962; Green, 1952; Horst, 1936, 1937; Horst & MacEwan, 1957; Hotelling, 1933a, 1933b; Howe, 1955; Hurley & Cattell, 1962; Johnson, 1964; Jöreskog, 1963; Mosier, 1939; Pearson, 1901; Schönemann, 1966; Tucker, 1951), and also for the formulation of side conditions in optimization problems, as will be shown in Section 5. Traces, it is true, could have been handled with the technique developed by Dwyer and MacPhail without much difficulty. As indicated above, this is not true for determinants, a second class of scalar functions which figure prominently in applied, especially multivariate, work (Anderson, 1958; Bargmann, 1957; Bargmann & Mah, 1961; Bock & Bargmann, 1966; Cooley & Lohnes, 1962; Dwyer, 1958; Graybill, 1961; Jöreskog, 1963; Kendall, 1957; McKeon, 1962; Rao, 1962; Todd, 1962). As will be shown the derivatives of these more complicated functions can be expressed in terms of those of traces.

In Section 2 the formal differentiation of traces will be discussed in some detail. A number of specific derivatives will be given which will serve as building blocks for later use as illustrated in Section 3, where a number of rules will be given for obtaining matrix derivatives of more complicated functions in terms of simpler ones. In Section 4 the connection with determinants will be made and in Section 5 a number of applied problems, most of which are well-known, will be reformulated in terms of the techniques discussed in the earlier sections.

2. *Formal Differentiation of Traces*[2]

Dwyer and MacPhail, in their paper, distinguish between two kinds of "matrix derivatives", (I) The derivative of a matrix with

[2]Parts of this section are taken from Schönemann (1964).

respect to a scalar, and (II) The derivative of a scalar with respect to a matrix. A special case of the first type would be the derivative of a vector function, which varies with some scalar variable, e.g., time. For example, a row vector $v' = \partial u'/\partial t = (0, 1, 2t)$ is obtained by differentiation the row vector $u' = u'(t) = (1, t, t^2)$ with respect to the independent scalar variable t. However for the present purposes the derivatives of type II are of more interest. A special case would be the derivative of a quadratic form $x'Ax$ with respect to the column vector x. Let $x' = (x_1, x_2, \ldots, x_i, \ldots, x_n)$ be a row vector of order n. Then the "formal derivative" of the scalar

$$f = x'Ax,$$

to be written $\partial f/\partial x$, will be defined as the column vector of partial derivatives $\partial f/\partial x_i$, i.e.,

[1]
$$\partial f/\partial x = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_i \\ \partial f/\partial x_n \end{bmatrix}$$

$$= (A + A')x, \text{ in this example.}$$

This derivative notation is often used in least-squares work (Aitken, 1959; Anderson, 1958; Bargmann, 1957; Bargmann & Mah, 1961; Graybill, 1961; Kaiser & Dickman, 1959; McKeon, 1962). Graybill (1961) also considers the fairly obvious generalization of the above definition to include matrices of independent variables, which is of major concern here.

When X is an $n \times k$ matrix, then $Y = X'AX$ is no longer a scalar. Suppose one considers the sum of the diagonal elements of Y. This sum (which is defined only for square matrices) is often called the "trace" of Y and denoted tr(Y), i.e., if

$$Y = (y_{ij}),$$

then

$$\text{tr } (Y) = \sum_i y_{ii}.$$

This trace is a scalar variable which can be treated by ordinary methods of calculus, if the partial derivatives with respect to each element x_{ij} in x are desired. But rather than considering each of these

partials separately, it may be convenient, for some purposes, to arrange them in a matrix of the same order as x, the independent (matrix) variable, and to denote this $n \times k$ matrix simply as $\partial \mathrm{tr}(Y)/\partial x$, i.e.,

$$[2] \qquad \partial \mathrm{tr}(X'AX)/\partial X = (\partial \mathrm{tr}(X'AX)/\partial x_{ij}),\ i=1, n, j=1, k.$$

So far few gains seem to derive from such a convention. One notes however, that the scalar $\mathrm{tr}(Y)$ is of a somewhat peculiar nature and subject to special algebraic conveniences which, if combined with the suggested notation, are apt to lead to considerable algebraic simplification in least-squares work.

To see this, one first notes that the trace of a matrix is but one member of a whole class of scalar functions of a square matrix, which are sometimes (Bellman, 1960, p. 95) denoted Φ_k, and which appear as coefficients in the characteristic polynomial of the matrix, say Y. Let Y be $n \times n$ with roots r_i, then

$$\Phi(r) = \Sigma_k\ (-1)^k\ \Phi_k\ r^{n-k} \qquad (k = 0,1, \ldots, n)$$

for all roots r, where Φ_n is the determinant of Y, Φ_1 is the trace of Y, and all Φ_k are chosen so that $\Phi_0 = 1$ (see, e.g., Bellman, 1960; Finkbeiner, 1960; Hohn, 1958).

For all Φ_k, the following property holds: Let $Y = ABC$ (A, B, C, all square). Then

$$\Phi_k\ (ABC) = \Phi_k\ (CAB) = \Phi_k\ (BCA),$$

but

$$\Phi_k\ (ABC) \neq \Phi_k\ (ACB),$$

in general. That is to say the scalar functions Φ_k are invariant under cyclic permutations of the fractors of Y. A special case, of course, is the commutativity for traces and determinants of matrix products with two factors, which holds for traces even where the factors are rectangular.

Furthermore traces, being based on square though not necessarily symmetric matrices, have the convenient property of invariance under transposition, i.e.,

$$\mathrm{tr}(Y) = \mathrm{tr}(Y'),$$

which is obvious because transposition does not affect the diagonal of a matrix.

These two properties can be used to partition the dependent variables of derivatives of traces into equivalence classes. For exam-

ple, $\text{tr}(XA) = \text{tr}(AX)$ by cyclic permutation, $\text{tr}(A'X') = \text{tr}(X'A')$ by transposition, so that all four "trace equivalent forms" i.e., the traces of XA, AX, $A'X'$, and $X'A'$ will have the same "formal derivative," A'.

On the other hand, it is a direct consequence of the definition in Equation 2 that transposition of the independent variable, i.e., X, will transpose the derivative, i.e.,

$$\partial \text{tr}(Y)/\partial X' = (\partial \text{tr}(Y)/\partial X)'.$$

The problem, then, is to find the derivatives of a representative of a trace-equivalence class. This can be accomplished in several ways, for example by recourse to summation notation, by recourse to "vector derivatives," or even by writing some of the terms of the trace explicitly and going back to matrices after some of the partial derivatives have been taken. To illustrate the vector technique, let

$$A = \begin{bmatrix} a_1' \\ a_i' \\ a_k' \end{bmatrix} \text{ and } X = [x_1, x_i, x_k]$$

where $k = m$. Then

$$\text{tr}\,(AX) = \sum_i a_i' X_i.$$

One verifies immediately that $\partial a_i' x_i / \partial x_i = a_i$, so that

$$\partial \text{tr}(AX)/\partial X = (a_1, \ldots, a_k) = A'.$$

In similar fashion a number of other derivatives can be found, some of which are summarized in Table 1.

Although it is not anticipated that this presentation will be entirely satisfactory for some of the more mathematically inclined readers it should be noted that the proposed technique is capable of further formalization, should that be desired. The following few remarks are intended to sketch a possible basis for such formalization.

First one notes that what has been called "symbolic vector differentiation" can be put on a somewhat more formal basis. Given a unit length vector w (i.e., $w'w = 1$) and some second vector x_0, consider the scalar

Table 1

Some Matrix Derivatives of Traces

Y	Trace-Equivalent Forms					Trace Derivatives	
	Cycl.	Permut.	Transpos.			trY/∂X	∂trY/∂X'
A		-	A'			0	0
X		-	X'			I	I
AX	XA		X'A'	A'X'		A'	A
A'X	XA'		X'A	AX'		A	A'
X'AX	XX'A	AXX'	X'A'X	A'XX'	XX'A	(A+A')X	X'(A+A')
XAX'	X'XA	AX'X	XA'X'	A'X'X	X'XA'	X(A+A')	(A'+A)X'

[3] $$d_w f(x_0) = \lim_{t \to 0} \frac{f(x_0 + tw) - f(x_0)}{t}$$

where t is a scalar. This limit, as is well known (see e.g., Crowell, 1962, p. 234; Faddeev & Faddeeva, 1963, p. 116), defines the "directional derivative" of f at the point x_0 "in the direction of w". If one chooses w successively so as to give the n directions parallel to each of the (mutually orthogonal) coodinate axes, i.e., if one lets $w^{(i)\prime} = (0, 0, \ldots, 1, \ldots, 0)$ (with unity as the ith component) then Equation 3 will yield $\partial f/\partial x_i$, for each i, in accordance with the standard definition of a partial derivative (see e.g., Olmstead, 1959, p. 366). Therefore, if one writes Equation 3 as the scalar product of some unknown vector u with the directional vector w

[4] $$D_w f(x_0) = w'u$$

then this argument shows that u is identical with what was called the "symbolic vector derivative" i.e.,

$$u = \partial f/\partial x,$$

evaluated at x_0 (since multiplication with $w^{(i)}$ simply picks off the ith component of u). To demonstrate that such a definition of $\partial f/\partial x$ indeed "works", consider again Equation 1:

$$f = x'Ax$$

$$D_w f(x) = \lim_{t \to 0} \{(x + tw)'A(X + tw) - x'Ax\}/t$$

$$= \lim_{t \to 0} (x'Ax + tw'Ax + tx'Aw + t^2\, w'Aw - x'Ax)/t$$

$$= \lim_{t \to 0} w'(A + A')x + tw'Aw$$

$$= w'(A + A')x,$$

whence

$$u = \partial x'Ax/\partial x = (A + A')x,$$

as in Table 1 (some authors, e.g., Rao, 1952, p. 21 ff. simply give $2Ax$ for this derivative, tacitly assuming that A is symmetric).

This approach is easily extended to the case where f is a trace and the independent variable a (not necessarily square) matrix. To see this one notes that matrices of a given order form a vector space, in the abstract sense. To see this also in the less abstract sense, suppose one "expands" a given matrix into a supervector having the columns of the matrix as components, for example. The expressions of the type $\mathrm{tr}(A'B)$ will correspond to scalar products of such supervectors. In particular, the so called "Frobenius norm" of a real matrix A, given by $F(A) = [\mathrm{tr}(A'A)]^{1/2}$ (Taylor, 1955, p. 227), corresponds to the "length" of such supervectors and can be used to measure the distance of the matrices from null. Making use of such a correspondence in Equations 3 and 4, one arrives at

$$D_Y f(X_0) = \lim_{t \to 0} \frac{f(X_0 + tY) - f(X_0)}{t},$$

and

$$D_Y f(X_0) = \mathrm{tr}Y'U, \text{ where now } \mathrm{tr}(Y'Y) = 1$$

and where the matrix U corresponds to what is called here the "formal matrix derivative of f" (f being a trace), i.e.,

$$U = \partial f/\partial X, \text{ evaluated "at the point" } X_0.$$

For illustration, consider $\partial f/\partial X$ where $f = \mathrm{tr}(X'AXB)$.

Here one finds for

$$D_Y f(X_0) = \lim_{t \to 0} \{\mathrm{tr}((X + tY)'A(X + tY)B - X'AXB)\}/t$$

$$= \lim_{t \to 0} \mathrm{tr}(X'AXB + tX'AYB + tY'AXB$$

$$+ t^2\, Y'AYB - X'AXB)/t$$

$$= \text{tr}\{Y'(A'XB' + AXB)\},$$

making use of

$$\text{tr}(X'AYB) = \text{tr}(BX'AY) = \text{tr}(Y'A'XB')$$

so as to be able to factor out Y'. The matrix postmultiplying Y' is the desire "formal derivative" of f, viz.

$$[5] \qquad \partial\text{tr}(X'AXB)/\partial X = AXB + A'XB'$$

3. Some General Rules for the Formal Differentiation of Traces

So far only two specific formulae for the formal differentiation of traces have been established:

$$[6] \qquad \partial\text{tr}AX/\partial X = A', \text{ and}$$

$$[7] \qquad \partial\text{tr}X'AXB/\partial X = AXB + A'XB'.$$

But these two specific formulae cover a large variety of possible matrix expressions if used in conjunction with a few general rules for manipulating such expressions. Of such general rules, three have been mentioned already:

$$[8] \qquad \partial\text{tr}Y/\partial X = \partial\text{tr}Y'/\partial X \text{ (transposition of dependent variable)}$$

$$[9] \qquad \partial\text{tr}UVW/\partial X = \partial\text{tr}WUV/\partial X \text{ (cyclic permutation)}$$

$$[10] \qquad \partial\text{tr}Y/\partial X' = (\partial\text{tr}Y/\partial X)' \text{ (transposition of independent variable).}$$

To illustrate the use of these rules, suppose one wishes to find the derivative

$$[11] \qquad \partial\text{tr}AX'/\partial X.$$

This expression is not obtainable from Equation 6 by transposition and/or cyclic permutation. Rather, it belongs to the second possible equivalence class for traces of two factors. Use of Equation 10 and a simple substitution yields a representative of this other equivalence class at once: From Equations 6, 8, and 9 one knows that

$$\partial\text{tr}BX/\partial X = \partial\text{tr}X'B'/\partial X = \partial\text{tr}B'X'/\partial X = B.$$

Setting $A = B'$ one obtains

$$\partial\text{tr}AX'/\partial X = A$$

as the desired result.

Likewise, one obtains for the corresponding alternative of Equation 7

$$\partial \text{tr}\, XAX'B/\partial X = BXA + B'XA'$$

as the reader may verify without difficulty. As will be seen in Section 5, expressions of this type are sometimes useful for formulating side conditions with the aid of Lagrange multipliers.

For most practical applications, it is hoped, the summary in the Appendix will serve to give the needed derivatives. But, as users of tables of integrals will appreciate, it is sometimes necessary to make some minor transformation to a more common form before tables can be applied. For this purpose the invariance under transposition and under cyclic permutations will be more convenient, as will be the "product rule," now to be discussed.

For its derivation a few notational conventions are needed. First, let $\partial Y/\partial t$ denote a matrix with elements $\partial y_{ij}/\partial t$. With this notation (which is rather common, see Bellman, 1960; Browne, 1958; Finkbeiner, 1960) one verifies

$$\frac{\partial UV}{\partial t} = \frac{\partial U}{\partial t} V + U \frac{\partial V}{\partial t} \tag{12}$$

as an immediate consequence of the product rule of elementary calculus and the definition of a matrix product. Similarly one finds (Finkbeiner, 1960)

$$\text{tr}(\partial Y/\partial t) = \partial(\text{tr}Y)/\partial t, \tag{13}$$

as an immediate consequence of the definition of $\partial Y/\partial t$, above, and that of the trace operator.

Now let it be agreed always to subscript the symbol for a variable (here the last letters of the alphabet) with "c" if they are to be held constant for purposes of differentiation. To illustrate, take $\partial \text{tr}U'AU/\partial U$, which is equal to $(A + A')U$ as a special case of Equation 7, with $B = I$. Subscripting U' with "c" would lead to $\partial \text{tr}U_c' AU/\partial U = (U'A)' = A'U$, as in Equation 6.

With this notation, and substituting a particular element x_{ij} (a scalar) for t, one can rewrite Equation 12 as

$$\partial \text{tr}(UV)/\partial x_{ij} = \partial(\text{tr}UV_c)/\partial x_{ij} + \partial(\text{tr}U_cV)/\partial x_{ij}$$

which, together with Equation 13, yields for a whole matrix X of such elements x_{ij}

$$\partial \text{tr}UV/\partial X = \partial \text{tr}UV_c/\partial X + \partial \text{tr}U_cV/\partial X \tag{14}$$

as the "product rule for formal differentiation of traces".

To illustrate its use, consider again Equation 5:

$$\begin{aligned}\partial \mathrm{tr}(X'AXB)/\partial X &= \partial \mathrm{tr}X'(AXB)_c/\partial X + \partial \mathrm{tr}X_c'XB/\partial X \\ &= \partial \mathrm{tr}X'(AXB)_c/\partial X + \partial \mathrm{tr}(BX'A)_cX/\partial X \\ &= AXB + A'XB'.\end{aligned}$$

In using this product rule, which evidently is of great convenience in practical work, care must be taken that the independent variable (X and/or X') appears separately as an independent variable as often on the right side of Equation 14 as it appears on the left side. To illustrate, the foregoing example was broken down into $\mathrm{tr}(X')(AXB)$. It could also have been broken down into $\mathrm{tr}(X'A)(XB)$ or even into $\mathrm{tr}\,(X')(A)(X)(B)$, although the latter partition would seem somewhat redundant. But all these partitions would have yielded the same, correct, result. In contrast, the partition $\mathrm{tr}(X'AX)(B)$ would yield an incorrect result as is easily verified.

As a further application of the product rule, consider expressions of the type $\mathrm{tr}(X^{-1}A)$, which one encounters in maximum likelihood problems involving the multivariate normal distribution, for example. The product rule can be put to use for finding the formal derivative of such expressions involving inverses:

Consider

$$\begin{aligned}\partial \mathrm{tr}Y^{-1}/\partial X &= \partial \mathrm{tr}Y^{-2}Y/\partial X \\ &= \partial \mathrm{tr}Y_c^{-2}Y/\partial X + \partial \mathrm{tr}Y^{-2}Y_c/\partial X,\end{aligned}$$

by the product rule. For the second member of the right side one finds

$$\begin{aligned}\partial \mathrm{tr}Y^{-2}Y_c/\partial X &= \partial \mathrm{tr}Y_c^{-1}Y^{-1}Y_c/\partial X + \partial \mathrm{tr}Y^{-1}Y_c^{-1}Y_c/\partial X \\ &= \partial \mathrm{tr}(Y_cY_c^{-1})Y^{-1}/\partial X + \partial \mathrm{tr}Y^{-1}(Y_c^{-1}Y_c)/\partial X \\ &= 2\partial \mathrm{tr}Y^{-1}/\partial X, \text{ whence} \\ \partial \mathrm{tr}Y^{-1}/\partial X &= \partial \mathrm{tr}Y_c^{-2}/\partial X + 2\ \partial \mathrm{tr}Y^{-1}/\partial X \text{ or} \\ \partial \mathrm{tr}Y^{-1}/\partial X &= -\partial \mathrm{tr}Y_c^{-2}Y/\partial X. \qquad [15]\end{aligned}$$

To generalize Equation 15 to a form more directly applicable to problems involving the multivariate normal distribution, let $U^{-1} = Y^{-1}A$ so that $U^{-2} = Y^{-1}AY^{-1}A$ and $U = A^{-1}Y$, where, of course, it is assumed that all these inverses exist. Then from Equation 15, $\partial \mathrm{tr}U^{-1}/\partial X + -\partial \mathrm{tr}U_c^{-2}U/\partial X = -\partial \mathrm{tr}(Y^{-1}AY^{-1})_cA^{-1}Y/\partial X$, so that $\partial \mathrm{tr}Y^{-1}A/\partial X = -\partial \mathrm{tr}(Y^{-1}AY^{-1})_cY/\partial X$, which contains Equation 15 as a special case.

Finally, two rules will be given which apply when the independent variable is restricted in some fashion. The most common restriction, for square independent variables would be symmetry. This particular restraint, however, is more conveniently handled with the help of Lagrange mutlipliers; i.e., by introducing the side condition $X - X' = 0$, which will be taken up in Section 5.1 and 5.2. At this point two other kinds of restrictions on X will be discussed viz. the case where

(i.) $X = D$ = diagonal, as, for example, the matrix U^2 in factor analysis, and the case where

(ii.) $X = xI$, i.e., X is a scalar matrix.

In the first case, where $X = D_x$ = diagonal, let $D_x = U$, i.e.,

$$\begin{bmatrix} x_1 & & 0 \\ & x_2 & \\ 0 & & x_n \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ u_{n1} & u_{n2} & \cdots & u_{nn} \end{bmatrix}$$

define a correspondence between the (unrestricted) matrix $U = (u_{ij})$ and the (diagonal) matrix D_x. Then the partial derivatives of some function f, with respect to the x_i can be expressed in terms of the partial derivatives of f with respect to the u_{ij} by means of the chain rule (for several variables as follows:

$$\frac{\partial f}{\partial x_1} = \frac{\partial f}{\partial u_{11}}\frac{\partial u_{11}}{\partial x_1} + \frac{\partial f}{\partial u_{12}}\frac{\partial u_{12}}{\partial x_1} + \cdots + \frac{\partial f}{\partial u_{nn}}\frac{\partial u_{nn}}{\partial x_1} = \frac{\partial f}{\partial u_{11}}$$

because $\partial u_{ij}/\partial x_1 = 0$ for all $j \neq 1$,

$= 1$ for all $j = 1$.

More generally, for any i,

$$\partial f/\partial x_i = \partial f/\partial u_{ii},$$

for analogous reasons, whence, for the case where $X + D_x$ = diagonal,

$$\partial f/\partial X = \text{diagonal } (\partial f/\partial U)$$

where U is unrestricted.

Precisely the same kind of reasoning yields, for the case where $X = xI$ = scalar,

[16] $$\partial f/\partial x = \text{tr}(\partial f/\partial U)$$

where U is unrestricted.

4. *The Reduction of Formal Derivatives of Determinants to Those of Traces*

Mention of what is here called "formal differentiation" of determinants appears rather early in the literature, often simply labelled "the derivative of a determinant" (Aitken, 1959; Anderson, 1958; Turnbull, 1960; Turnbull & Aitken, 1932; Wedderburn, 1964). The reason, perhaps, is that such derivatives emerge quite naturally when one considers the expansion of a determinant by rows, or columns. Such an expansion, as is well known, reproduces the determinant as a weighted sum of the elements in the row (or column), the weights being the signed minors corresponding to these elements. These minors are known not to contain any element of the row (or column); in fact, one customarily obtains minors by "striking out" the row and column in which the element appears. This observation leads therefore to the conclusion that "the derivative of a determinant" is simply the transposed adjoint of the corresponding matrix, or

$$\partial|Y|/\partial Y = Q' \tag{17}$$

where

$$Q = |Y|Y^{-1},$$

and $|Y|$ denotes the determinant of Y.

Consider now the more useful case where Y is some matrix product containing the independent matrix variable X (and/or its transpose) one or more times and where the problem is to find $\partial|Y|/\partial X$. By the chain rule

$$\begin{aligned} \partial|Y|/\partial x_{pq} &= \sum_i \sum_j [\partial|Y|/\partial y_{ij}][\partial y_{ij}/\partial x_{pq}] \\ &= \sum_i \sum_j q_{ji} [\partial y_{ij}/\partial x_{pq}] \end{aligned} \tag{18}$$

where y_{ij} and q_{ij} are elements of Y and $Q = |Y|Y^{-1}$, respectively, x_{pq} is a fixed element of the matrix X and where use was made of the result in Equation 17. Equation 18 can also be written as a trace of a matrix product, the matrices being $Q' = \partial|Y|/\partial y_{ij}$, with elements q_{ij}, and $\partial Y/\partial X$ with elements $\partial y_{ij}/\partial x_{pq}$, for a fixed x_{pq}, as in Section 3. With this notation Equation 18 could also be written

$$\begin{aligned} \partial|Y|/\partial x_{pq} &= \mathrm{tr} Q_c \, \partial Y/\partial x_{pq} \\ &= \partial \mathrm{tr} Q_c Y/\partial x_{pq}, \end{aligned} \tag{19}$$

where again the fact $\mathrm{tr}(\partial Y/\partial t) = \partial(\mathrm{tr}Y/\partial t)$ was used, as in the derivation of the product rule. The form in Equation 19 allows for immediate generalization to a whole matrix of elements x_{pq}, giving

$$\partial|Y|/\partial X = \partial \mathrm{tr} Q_c Y/\partial X \qquad [20]$$

as the desired result.[3] Formula (20) then allows the reformulation of problems involving determinants into problems which involve only traces, the latter being much more tractable functions. A brief example will suffice for illustration:

$$\partial|AX|/\partial X = \partial \mathrm{tr} Q_c AX/\partial X = (Q_c A)' = A'Q_c'$$
$$= |AX|A'(AX)^{-1'}.$$

5. *Some Illustrative Examples*

5.1 Lagrange Multipliers

Seldom if ever will the user be confronted with an optimization problem where the unknown matrix X is entirely unrestricted. It may be appropriate, therefore, to preface the discussion of some specific applications of the proposed technique with a brief sketch of the "method of Lagrange multipliers". This method, which is employed quite frequently when certain (side) conditions are to be imposed on the solutions x_{ij} in X, can be formulated in terms of traces.

To impose restraints on the solution matrix (or vector, as the case may be) often but not always amounts to searching for a lesser than the "best" optimum of the function (boundary points ignored), namely an optimum attained in the subspace defined by the constraints. Sometimes this subspace contains the "best" optimum. In such cases a constraint may be intended to select a particular solution out of a subset of equivalent solutions all which optimize the function. This is the case, for example, in certain scaling problems, when origin and overall dispersion are arbitrary, or in certain least-squares problems where a given matrix is to be approximated by the Gram product (XX') of an unknown matrix X, which then is only defined up to rotation. In this sense the well-known property of principal components to be uncorrelated in the sample is simply a question of identification, not a direct consequence of a least squares formulation. The same problem arises in factor analysis, when the stipulation is made that, e.g.,

[3]The foregoing line of argument was suggested by Professor Bock. It is better than the present author's original proof of Equation 20, which is therefore omitted.

$F'U^{-2}F$ be diagonal, as in Bargmann (1957) and Bargmann and Mah (1961), for example. In all these cases the constraint may be disregarded for purposes of optimization with the result that the conditional equations will have more than one solution, all yielding the "best" optimum of the function. Having arrived at this set of solutions we then invoke the constraint to select the desired solution (for an example see Section 5.8).

There are other cases where the constraint genuinely affects the optimum itself. For example, if it is desired to find a transformation matrix T so that AT approximates a given B in a least-squares sense, then the solution will, in general, be "best" if T is left entirely unspecified, as in regression problems (Graybill, 1961). It will be somewhat "less good" in general, if the relatively weak condition $\text{diag}(T'T) = I$ is imposed, as in the "oblique Procrustes problem" (Hurley & Cattell, 1962; Mosier, 1939) and will be "least good" (i.e. the sum of the squares of discrepancies will be largest), in general, if the stronger restraint is imposed that $T'T = I$, as in the "orthogonal Procrustes problem" (Green, 1952; Schönemann, 1966). In such cases the preferred method of introducing one or more side conditions is the technique of Lagrange multipliers (see, e.g., Taylor, 1955, p. 198) which, perhaps, owes some of its popularity to the fact that it applies whether or not some of the unknowns can be eliminated algebraically. Evidently such an elimination becomes more difficult with more unknowns.

The method is described most conveniently in actual use with a small example. Let a half-sphere be described by

$$(x - a)^2 + (y - b)^2 + (z - c)^2 = k^2, z \geq 0$$

with $c = 0$. The function z evidently attains its absolute maximum k at the center of the circle, (a,b), in the xy-plane. Now suppose the side condition $x + y - m = 0$ is imposed. A direct solution of this new problem would be to eliminate one of the variables, say y, with the aid of the side condition, and to optimize the new problem with respect to one unknown, x:

$$y = m - x$$

$$f = z^2 = k^2 - \{(x - a)^2 + (m - x - b)^2\}$$

$$\partial z^2/\partial x = 0 - \{2(x - a) + 2(m - x - b)(-1)\} = 0$$

$$x = (m + a - b)/2$$

In contrast, to apply the method of Lagrange multipliers, an additional

(third) unknown (viz. the Lagrange multiplier) is introduced, let it be u. A new function, say g, is set up where g is the sum of the original optimization criterion (f) and the product h being a formulation of the restraint such that, where it is satisfied, $h = 0$. Here

[21] $$g = f + uh = z^2 + (x + y - m).$$

This equation is to be differentiated with respect to x and y, yielding, in this case, two equations g_1, g_2, which together with h define a solution for the three unknowns x, y and u:

$$g_1 = \partial g/\partial x = -2(x - a) + u = 0$$

$$g_2 = \partial g/\partial y = -2(y - b) + u = 0$$

$$h = x + y - m = 0$$

whence

$$x - a = u/2 = y - b, x = a + y - b = a + m - x - b$$

$$x = (a + m - b)/2,$$

as before.

In the more general case where m side conditions are to be imposed on the solution, m such Lagrange multipliers will be needed, i.e., in this case

$$g = f + \sum_{i=1}^{m} u_i h_i (x_1, x_2, ..., x_n)$$

will have to be differentiated partially with respect to the n unknowns x_i and the resulting $n + m$ equations obtained by setting these partial derivatives to zero need be solved for the n x_i and the m u_i.

This method is used widely in the statistical and psychometric literature. (e.g. Anderson, 1958; Cliff, 1966; Cooley & Lohnes, 1962; Dwyer, 1958; Edgerton & Kolbe, 1936; Graybill, 1961; Green, 1952; Horst, 1937; Hotelling, 1933, 1935a; Joreskog, 1963; Kendall, 1957; McKeon, 1962; Mosier, 1939; Rao, 1952; Schönemann, 1965, 1966; Tucker, 1951), and in view of this popularity it is rather fortunate that it can be reformulated in terms of traces.

To see this it will suffice to consider expressions of the type

[22] $$\mathrm{tr}A'B = \sum_i \sum_j a_{ij} b_{ij}$$

This is a sum of products of "corresponding" elements in A and B, with the element a_{ij} corresponding to the element b_{ij}. Suppose now the

object is to impose a set of side conditions on the $n \times m$ elements x_{ij} of an unknown matrix X, and suppose these side conditions can be formulated in terms of matrix algebra involving the matrix X and some constant matrix C. Two (quite arbitrary) examples might be

$$X'A + C \text{ or } XPX' \; AC = C$$

Such conditions can be reformulated as

$$X'A - C = 0 \text{ or } XPX' + AX - C = 0$$

Let the resulting matrix, which is equal to the null matrix, be called $H = H(X)$, with elements h_{ij}, and let H be of order $p \times q$. The method of Lagrange multipliers requires that every one of the $p \times q$ elements h_{ij} be multiplied by an unknown multiplier u_{ij} and all these $p \times q$ products be summed before adding them to the scalar function $f(X)$ to obtain g(X) as in Equation 21. In view of Equation 22 this end is accomplished by simply writing

$$g = f + \text{tr}U'H. \tag{23}$$

Assuming that $f(x)$ was a scalar function which allowed expression in terms of traces and/or determinants involving X, Equation 23 could then be treated by the techniques described in the preceding pages, as illustrated in the following pages.

5.2 Least-squares Problem 1: To Approximate a Given Matrix A by a Symmetric Matrix X

In this, as in all other linear least-squares problems, the object is to minimize the sum of squares of a matrix of (possible weighted) discrepancies. This sum of squares, fortunately, is also expressible in terms of traces. In the specific example, let

$$A = X + E.$$

The object is to find X subject to $H(X) = X - X' = 0$ so as to minimize $f = \text{tr}E'E$. Therefore, one has to differentiate (Appendix, F25)

$$g = \text{tr}E'E + \text{tr}U'(X - X').$$

This derivative, according to the Appendix (F11, F12, F28) comes to

$$\partial g/\partial X = -2A + 2X + U' - U$$

and is to be set to zero as a necessary condition of an extremum of g. Hence

$$X = A + (U - U')/2$$

$$X' = A' + (U' - U)/2$$

or

$$X = (A + A')/2$$

5.3 Least-Squares Problem 2: To Approximate a Given Matrix A by an Orthogonal Matrix X

This problem was considered by Gibson (1962), using a somewhat different route. It differs from the preceding one in terms of the side condition. Let again

$$A = X + E.$$

Here the object is to find X subject to $H(X) = X'X - I = 0$ so as to minimize $f = \text{tr}E'E$. Therefore, one has to differentiate

$$g = \text{tr}E'E + \text{tr}U(X'X - I).$$

This derivative, according to the Appendix (F11, F12, F14, F21, F28), comes to

$$\partial g/\partial X = -2A + 2X + X(U + U').$$

$\partial g/\partial X = 0$ leads to

$$X(U + U')/2 = A - X, \text{ or, in view of } X'X = I$$

$$(U + U')/2 = X'A - I = (U + U')'/2, \text{ so that}$$

$$X'A = A'X.$$

This use of the symmetry of $U + U'$ to eliminate the (unknown) matrix of Lagrange multipliers U is often useful in problems of this type (Schönemann, 1964, 1966). Assuming now A to have an Eckart-Young decomposition (Eckart & Young, 1938)

$$A = VDW'$$

(with $W'W = V'V = Z$, D = diagonal), one finally arrives at

$$X = VW',$$

as shown in Schönemann (1964, 1966).

5.4 Least-Squares Problem 3. A Factor Model Proposed by Jöreskog

Jöreskog (1963), in his monograph, presents a new factor model which is closely allied with those of Guttman, Harris, Kaiser, Lawley

and Rao (Bellman, 1961; Harman, 1960; Howe, 1955; Kendall, 1957; Lawley, 1940; Lawley & Maxwell, 1963; Rao, 1955), which allows for (appropriate) statistical tests, and which does not require iteration. Jöreskog arrives at basically the same technique from various directions. On p. 36 he treats the problem in a least-squares sense. This treatment will be taken up here in a purely technical manner. For a statement of the logic and foundation of the Jöreskog model the reader is directed to Jöreskog (1963).

With a minor change in notation (we prefer to write D^2 where Jöreskog writes D, to avoid broken exponents, and we use F where he uses Λ) the problem is to minimize the (weighted) sum of squares of

$$E = S - FF' - tD^{-2}$$

where the weights are given by $D = -(\text{diag}(S^{-1}))^{1/2}$ and D is assumed to be known. The optimization is under choice of F and the scalar t. The criterion Jöreskog wants to minimize is, in the present notation,

$$u = \text{tr}(E^{*\prime}E^{*})$$

where

$$E^{*} = DED = D(S - FF')D - tI,$$

so that

$$u = \text{tr}(DSD^2SD + t^2I - 2tDSD - 2DSD^2\,FF'D + 2tDFF'D + DFF'D^2FF'D)$$

making use of the invariance of traces under cyclic permutation, where convenient. Use of the Appendix (F12, F13, F24) gives for the formal derivative with respect to F

$$\partial u/\partial F = -4D^2\,SD^2F + 4tD^2\,F + 4D^2FF'D^2F,$$

to be set to zero. Multiplication by $-D^2/2$ leads to conditional equation

[25] $$SD^2F - tF - FF'D^2F = 0,$$

which is Jöreskog's Equation 7.16, but for notation. To differentiate u with respect to the scalar t one could use the rule in Equation 16, rewriting Equation 24 as

$$u_T = \text{tr}(\text{const} + T'T - 2TDSD + 2TDFF'D)$$

with T now unrestricted. Again by the Appendix (F12, F13)

$$\partial u_T/\partial T = 2T - 2DSD + DFF'D$$

whence

[26] $$\partial u/\partial t = 2\text{tr}(tI - DSD + DFF'D)$$

which is Jöreskog's Equation 7.15. Equations 25 and 26 allow for an algebraic (e.g. noniterative) solution of F and t, as shown by Jöreskog (1963).

5.5 Maximum Likelihood Problem 1: To Estimate the Covariance Matrix of the Multivariate Normal Distribution

This problem is sometimes (Anderson, 1958, p. 44) solved by use of a theorem which, in effect, states that under certain mild conditions a maximum likelihood estimate of a one-one transformation of a set of parameters is given by the same transformation of the maximum likelihood estimates of the parameters (see e.g., Anderson, 1958, p. 48). This theorem allows us to phrase the maximization problem in terms of Σ^{-1}, rather than Σ. But this indirect route is not necessary because use of F16 in the Appendix leads to quite simple algebra even for the direct solution.

The logarithm of the likelihood function of the multivariate normal distribution is given by

$$\ln L = \text{const} - N/2\{\ln|\Sigma| + \text{tr}\Sigma^{-1} S/N + (\bar{x} - \hat{\mu})'\Sigma^{-1}(\bar{x} - \hat{\mu})\},$$

(see, e.g., Anderson, 1958, p. 46 or Jöreskog, 1963, p. 34), where

$$S = 1/N \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})'$$

and x, $\bar{x}$, and $\hat{\mu}$ are $p \times 1$. At the point $\bar{x} = \hat{\mu}$ this reduces to

[27] $$\ln L = c_1 + c_2[\ln|\Sigma| + \text{tr}(\Sigma^{-1}S)].$$

Using $Q = |\Sigma|\Sigma^{-1}$, as in Section 4 and the Appendix (F12, F16, F29) one finds for the formal derivative of $\ln L$

$$\begin{aligned}\partial \ln L/\partial\Sigma &= c_2\,[(|\Sigma_c|^{-1}\,\partial\text{tr}Q_c\Sigma/\partial\Sigma) + \partial\text{tr}(\Sigma^{-1}S\Sigma^{-1})_c\Sigma/\partial\Sigma] \\ &= c_2\,(\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}),\end{aligned}$$

to be set to zero as a necessary condition for an extremum. Upon pre- and postmultiplication with Σ

$$\hat{\Sigma} = S.$$

5.6 Maximum Likelihood Problem 2. To Estimate the Components of a Specified Structure of the Covariance Matrix

This problem has been treated by Bock and Bargmann (1966). In

its more general form the problem is to estimate the matrices X, Y, and Z, given S, in

[28] $$\Sigma = Z'\,(KXK' + Y)Z,$$

allowing for the options that Z = diagonal, X = diagonal, Y = diagonal, $Y = yI$ = scalar. Differentiating Equation 27

$$\ln L = c_1 + c_2\,(\ln|\Sigma| + \mathrm{tr}\Sigma^{-1}S)$$

with respect to any of the unknown matrices, say V (generically), one obtains from the Appendix (F16, F29)

[29] $$\begin{aligned}\partial \ln L/\partial V &= c_2\,(|\Sigma_c^{-1}|\partial \mathrm{tr}Q_c\Sigma/\partial V - \partial \mathrm{tr}(\Sigma^{-1}S\Sigma^{-1})_c\Sigma/\partial V\\ &= c_2\,(\partial \mathrm{tr}(\Sigma^{-1} - \Sigma^{-1}\,S\Sigma^{-1})_c\Sigma/\partial V\\ &= c_2\,\partial \mathrm{tr}B\Sigma/\partial V\end{aligned}$$

where

[30] $$B = (\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})_c = \Sigma_c^{-1}(\Sigma - S)_c\Sigma_c^{-1}$$

and Σ is defined, according to the model, as in Equation 28. This approach then reduces the seemingly rather involved problem to a quite simple one. Some solutions are:
Upon differentiation with respect to X, X unrestrained

[31] $$\partial \ln L/\partial X = K'Z(\Sigma^{-1}(\Sigma - S)\Sigma^{-1})Z'K = W, \text{ say;}$$

upon differentiation with respect to X, X = diagonal (F26)

$$\partial \ln L/\partial X = \text{diagonal } (W)$$

where W is defined as in Equation 31.
Upon differentiation with respect to Y, Y unrestrained

[32] $$\partial \ln L/\partial Y = Z'(\Sigma^{-1}(\Sigma - S)\Sigma^{-1})Z = T, \text{ say;}$$

upon differentiation with respect to Y, Y = diagonal (F26)

$$\partial \ln L/\partial Y = \text{diagonal } (T),$$

and upon differentiation with respect to $Y = yI$ = scalar (F27)

$$\partial \ln L/\partial Y = \mathrm{tr}(T)$$

where T is defined, in both cases, as in Equation 32.

Finally, upon differentiation with respect to Z, Z unrestrained,

[33] $$\partial \ln L/\partial Z = 2(KXK' + Y)Z(\Sigma^{-1}(\Sigma - S)\Sigma^{-1}) = R, \text{ say;}$$

and upon differentiation with respect to Z, Z = diagonal (F26)

$$\partial \ln L/\partial Z = \text{diagonal } (R),$$

with R as in Equation 33.

Since all these solutions are given implicitly, some iterative scheme will generally be needed to obtain numerical answers. Bock and Bargmann have employed successfully the Newton-Raphson method for estimating the parameters in the above models when Z, X, and Y are diagonal (Bock and Bargmann, 1966). This method, however, requires knowledge of the second order derivative which can not be obtained by the methods in this paper. As an alternative the method of gradients (Taylor, 1955, p. 276) or some other scheme based only on first order derivatives may be tried. Convergence will be somewhat slower with the latter methods.

5.7 Maximum Likelihood Problem 3: Lawley's Solution for the Factor Model

In this model the postulated structure is given by

$$\Sigma = FF' + U^2$$

Lawley (e.g., in Rao, 1952, p. 10) derives estimates of F and U^2 by treating each separately as an independent variable. Thus, using Equations 29 and 30 for brevity, one obtains (F13) for

[34] $$\partial \ln L/\partial U^2 = \partial \text{tr} R(FF' + U^2)/\partial F = \partial \text{tr} BFF'/\partial F = 2BF,$$

and (F12) for

$$\partial \ln L/\partial U^2 = \partial \text{tr} B(FF' + U^2)/\partial U^2 = \partial \text{tr} BU^2/\partial U^2 = B$$

if U^2 were unrestricted, so that (F26)

$$\partial \ln L/\partial U^2 = \text{diagonal } (B)$$

since U^2 = diagonal. Whence, in view of the definition of B in Equation 30

$$BF = 0 \text{ or } (\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})F = 0$$
$$\text{and diag}(B) = 0 \text{ or diag } (\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}) = 0$$

give the conditions for $\ln L$ (Equation 27) to be stationary. These equations are given, for example, in Rao (1952). They can be shown to lead to an eigenproblem on $U^{-1}SU^{-1}$, which is also the end result of a somewhat different approach by Rao (1955).

5.8 Maximum Likelihood Problem 4: To Maximize a Likelihood Ratio Criterion in Factor Analysis

Bargmann (1957, p. 48) wishes to maximize a likelihood ratio criterion for a test of independence in a partial covariance matrix. The criterion, which has been used and discussed by various authors (especially Howe, 1955), can be given an intuitive interpretation as measuring the departure of a rescaled residual matrix from the identity matrix. That is, once all common factors have been partialled out of a covariance matrix the remaining partial covariances should be a sample from a population of uncorrelated variables so that this matrix, upon normalization, should be close to the identity. The criterion in this or some other manner arrived at is

$$u = |U^{-1}(R - FF')U^{-1}|,$$

where R is a correlation matrix (i.e., S normalized by rows and columns so that $\text{diag}(R) = I$), F is a pattern of uncorrelated common factors and U^2 is a diagonal matrix of covariance of the uncorrelated unique factors. The normalization of S into R has introduced a dependency between the unknowns F and U^2, viz., $U^2 = I - \text{diag}(FF')$. Therefore F and U^2 can no longer be treated as two independent matrix variables and the chain rule will have to be used. This, as will be seen, complicates matters slightly. For convenience let $u_1 = |U^2|$, $u_2 = |R - FF'|$ so that $u = u_1u_2$ and $\partial u/F = u_2\partial u_1/\partial F + u_1\partial u_2/\partial F$, by rules of elementary calculus. Of the two derivatives involved the first one, $\partial u_1/\partial F$, is the more difficult:

$$\partial u_1/\partial F = \partial|U^{-2}|^{-1}/\partial F = -|U_c^2|^{-2}\,\partial|U^2|/\partial F,$$

all by elementary rules. Now

$$\begin{aligned}\partial|U^2|/\partial F &= \partial|I - H^2|/\partial F = \partial|I - \text{diag}(FF')|/\partial F \\ &= \partial\text{tr}Q_c[I - \text{diag}(FF')]/\partial F = -\partial\text{tr}[Q_c\text{diag}(FF')]/\partial F\end{aligned}$$

seems to present a problem. But note that $Q_c = |U^2|U^{-2} = \text{diagonal}$ (as well as $\text{diag}(FF')$, of course) so that

$$\text{tr}(Q_c\,(\text{diag}(FF'))) = \text{tr}Q_c\,FF', \tag{37}$$

because only the diagonal elements are involved on the right side of Equation 37.
Hence (F13, F29)

$$\partial|U^2|/\partial F = -\partial\text{tr}Q_cFF'/\partial F = -2Q_cF = -2|U^2|U^{-2}F$$

which in conjunction with Equation 36 gives

[38] $$\partial u_1/\partial F = 2|U^2|^{-1}U^{-2}F.$$

The second derivative is quite simple in comparison:

$$\begin{aligned}\partial u_2/\partial F &= \partial|R - FF'|/\partial F = \partial \mathrm{tr} Q_c(R - FF')/\partial F \\ &= -\partial \mathrm{tr} Q_c FF'/\partial F = -2Q_c F \\ &= -2|R - FF'|(R - FF')^{-1}F.\end{aligned}$$

This result, together with Equations 38 and 35 combine to yield for the derivative of the criterion u

$$\begin{aligned}\partial u/\partial F &= (|R - FF'|\,|U^2|^{-1}\,U^{-2}F - |U^{-2}|\,|R - FF'|\,[R - FF']^{-1}F) \\ &= |R - FF'|\,|U^2|\,(U^{-2}F - [R - FF']^{-1}F)\end{aligned}$$

which is Bargmann's Equation (5.23), (1957, p. 49).

Premultiplying $\partial u/\partial F = 0$ by $R - FF'$ and selecting a solution F and U^2 for which $F'U^{-2}F$ is diagonal leads to

[39] $$(RU^2 - I)F = F(F'U^{-2}F)$$

which, in view of the identification condition, defines F and U^2 implicitly as an eigenproblem for $RU^{-2} - I$, or equivalently, for $U^{-1}(R - U^2)U^{-1}$ (since the Φ_k of Equation 2. are invariant under cyclic permutation of the factors, the roots must be). Whence it is seen that the Howe-Bargmann approach, which can be formulated independently of any distribution assumption, if desired, is equivalent to the Lawley-Rao approach (Bargmann & Mah, 1961; Howe, 1955) of Section 5.7 as Bargmann himself points out. He also notes, however, that Equation 39 is not necessarily the most efficient algorithm for computing F and U^2.

In most of these examples the problem of identification was ignored. It should be clear, however, that it may be the more difficult one, in a given context. Once the formal derivative $\partial f/\partial X$ has been found one deals with an algebraic problem of isolating X. This problem will not always have an easy solution so that iteration may become imperative. An algebraic solution, though, is much to be preferred, when it exists. It is contended that the search for such a solution is facilitated by having the conditions stated in matrix form, which is the primary purpose of formal matrix differentiation.

Appendix

A Summary of Some Results in Formal Matrix Differentiation

1. Specific Results:

(F11) $\partial \mathrm{tr} A/\partial X = 0$,

A, B, constants
X, U, V, W, dependent variables
Y independent variable

(F12) $\partial \mathrm{tr} AX/\partial X = A'$

(F13) $\partial \mathrm{tr} X'AX/\partial X = (A + A')X$

(F14) $\partial \mathrm{tr} X'AXB/\partial X = AXB + A'XB'$

(F15) $\partial \mathrm{tr} Y^{-1}/\partial X = -\partial \mathrm{tr} Y_c^{-2} Y/\partial X$

(F16) $\partial \mathrm{tr} Y^{-1}A = -\partial \mathrm{tr}(Y^{-1}AY^{-1})_c Y/\partial X$

2. General Rules for Obtaining other Derivatives:

(F21) $\partial f/\partial X' = (\partial f/\partial X)'$ — Transposition of independent variable

(F22) $\partial \mathrm{tr} UVW/\partial X = \partial \mathrm{tr} WUV/\partial X = \partial \mathrm{tr} VWU/\partial X$ — Invariance under cyclic permutation

(F23) $\partial \mathrm{tr} Y/\partial X' = \partial \mathrm{tr} Y/\partial X$ — Invariance under transposition of dependent variable

(F24) $\partial \mathrm{tr} UV/\partial X = \partial \mathrm{tr} U_c V/\partial X + \partial \mathrm{tr} UV_c/\partial X$ — Product rule

where a matrix carrying the subscript "c" is to be regarded as a constant for further differentiation.

3. Restraints on the Independent Matrix Variable X:

$X = X'$: include

(F25) $\mathrm{tr} U(X - X')$,

where U in an (unknown) matrix of Lagrange multipliers, differentiate and solve for U and X (see F28 below).

X = diagonal:

(F26) $\partial \mathrm{tr} Y/\partial X$ = diagonal $\{\partial \mathrm{tr} Y/\partial W\}$, where W is unrestricted.

$X = xI$ = scalar:

(F27) $\partial \mathrm{tr} Y/\partial X = \mathrm{tr}\{\partial \mathrm{tr} Y/\partial W\}$

where W is unrestricted.

4. Side Conditions:

To maximize f so that $A(X) = 0$ differentiate

(F28) $g = f + \mathrm{tr} U'A$

with respect to X and solve for X and the matrix of Lagrange multipliers U.

5. Determinants:

(F29) $\partial |Y|/\partial X = \partial \mathrm{tr} Q_c\, Y/\partial X$, where $Q = |Y| Y^{-1}$

References

Aitken, A. C. (1959). *Determinants and matrices.* Edinburgh: Oliver Boyd.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis.* New York: Wiley.

Bargmann, R. (1957). *A study of independence and dependence in multivariate normal analysis.* Chapel Hill, N.C.: Institute of Statistics, Mimeo Series #186.

Bargmann, R. E., & Mah, V. (1961). *Theory of least squares, part one.* Lecture Notes, Virginia Polytechnic Institute.

Bellman, R. (1960). *Introduction to matrix analysis.* New York: McGraw Hill.

Bellman, R. (1961). *Adaptive control processes: A guided tour.* Princeton: Princeton University Press.
Bock, R. D., & Bargmann, R. (1966). Analysis of covariance structures. *Psychometrika, 31,* 507–534.
Browne, E. T. (1958). *Introduction to the theory of determinants and matrices.* Chapel Hill, N.C.: University of North Carolina Press.
Cattell, R. B. (1944). Parallel proportional profiles and other principles for determining the choice of factors by rotation. *Psychometrika, 9,* 267–283.
Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika, 31,* 33–42.
Cooley, W. W., & Lohnes, P. R. (1962). *Multivariate procedures for the behavioral sciences.* New York: Wiley.
Crowell, R. H., & Williamson, R. E. (1962). *Calculus of vector functions.* Englewood Cliffs: Prentice Hall.
Dwyer, P. S., & MacPhail, M. S. (1948). Symbolic matrix derivatives. *Annals of Mathematical Statistics, 19,* 517–534.
Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1,* 211–218.
Edgerton, H. A., & Kolbe, L. E. (1936). The method of minimum variation for the combination of criteria. *Psychometrika, 1,* 183–188.
Edgerton, H. A., & Kolbe, L. E. (1936). The method of minimum variation for the combination of criteria. *Psychometrika, 1,* 183–188.
Faddeev, D. K., & Faddeeva, V. N. (1963). *Computational methods of linear algebra.* San Francisco: W. H. Freeman.
Finkbeiner, D. T. (1960). *Introduction to matrices and linear transformations.* San Francisco: W. H. Freeman.
Gantmacher, F. R. (1959). *The theory of matrices, Vol I and II.* New York: Chelsea.
Gibson, W. A. (1962). On the least-squares orthogonalization of an oblique transformation. *Psychometrika, 27,* 193–196.
Graybill, F. R. (1961). *An introduction to linear statistics. Vol 1.* New York: McGraw-Hill.
Green, B. F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika, 17,* 429–440.
Harman, H. H. (1960). *Modern factor analysis.* Chicago: University of Chicago Press.
Hohn, F. E. (1958). *Elementary matrix algebra.* New York: MacMillan.
Horst, P. H. (1936). Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika, 1,* 53–60.
Horst, P. H. (1937). A method of factor analysis by means of which all coordinates of the factor matrix are given simultaneously. *Psychometrika, 2,* 225–236.
Horst, P. H. (1963). *Matrix algebra for social scientists.* New York: Holt, Rinehart and Winston.
Horst, P. H., & MacEwan. (1957). Optimal test length for multiple prediction: the general case. *Psychometrika, 22,* 311–324.
Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417–441, 498–520.
Hotelling, H. (1935a). The most predictable criterion. *Journal of Educational Psychology, 26,* 139–142.
Hotelling, H. (1935b). Relations between two sets of variates. *Biometrika, 28,* 321–377.
Howe, W. G. (1955). Some contributions to factor analysis. USAEC Rep. ORNL-1919.
Hurley, J. R., & Cattell, R. B. (1962). Producing direct rotation to test a hypothesized factor structure. *Behavioral Science, 7,* 258–262.
Johnson, R. M. (1964). The minimal transformation to orthonormality. Paper presented at a joint meeting of the Psychonomic Society and the Psychometric Society, Niagra Falls.
Kaiser, H. F., & Dickman, K. W. (1959). Analytic determination of factors. *American Psychologist, 14,* 425–430.
Keller, J. B. (1962). Factorization of matrices by least squares. *Biometrika, 49,* 239–242.
Kendall, M. G. (1957). *A course in multivariate analysis.* New York: Hafner.
Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of Royal Society of Edinburgh,* A, *60,* 394–399.
Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as a statistical method.* London: Butterworths.

McKeon, J. J. (1962). Canonical analysis: some relations between canonical correlation, factor analysis, discriminant functional analysis and scaling theory. Technical Report, Urbana, Illinois: University of Illinois.

Mosier, C. I. (1939). Determining a simple structure when loadings for certain tests are known. *Psychometrika, 4,* 149–162.

Munroe, M. E. (1963). *Modern multidimensional calculus.* Reading: Addision-Wesley.

Nel, D. G. (1980). On matrix differentiation in statistics. *South African Statistical Journal, 14,* 137–193.

Onerig, E. D. (1963). *Linear algebra and matrix theory.* New York: Wiley.

Olmstead, J. M. H. (1959). *Real variables.* New York: Appleton-Century-Croft.

Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. *Philosophical Magazine, 2,* 6th series, 557–572.

Perlis, S. (1952). *Theory of matrices.* Cambridge: Addision-Wesley.

Rao, C. R. (1952). *Advanced statistical methods in biometric research.* New York: Wiley.

Schönemann, P. H. (1964). A solution of the orthogonal Procrustes problem with applications to orthogonal and oblique rotation. Unpublished doctoral dissertation, University of Illinois. O.N. 65-902. University Microfilms, Inc., Ann Arbor, Michigan.

Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika, 31,* 1–10.

Taylor, A. (1955). *Advanced calculus.* Boston: Ginn and Company.

Todd, J. (Ed.) (1962). *Survey of numerical analysis.* New York: McGraw-Hill.

Tucker, L. R. (1951). A method for synthesis of factor analysis studies. A. G. O. Personnel Research Section Report No. 984, Department of the Army.

Turnbull, H. W. (1960). *The theory of determinants, matrices and invariants.* New York: Dover.

Turnbull, H. W., & Aitken, A. C. (1932). *An introduction to the theory of canonical matrices.* London: Blackie and Son (Dover reprint 1961).

Wedderburn, J. H. M. (1964). *Lectures on matrices.* New York: Dover.

Wrobleski, W. J. (1963). Extension of the Dwyer-MacPhail matrix derivative calculus with applications to estimation problems involving errors-in-variables and errors-in-equations. Technical Report, The University of Michigan.