KM 9(1983)
pag 58 - 84

THE USE OF GENERALIZED INVERSES IN RESTRICTED MAXIMUM LIKELIHOOD

F.J. Henk Don
Central Planning Bureau, Van Stolkweg 14
2585 JR The Hague - Netherlands

Abstract:

This paper applies the calculus of generalized inverses and matrix differentials to restricted maximum likelihood estimation. Some calculus results are derived for the first time. The ML estimation of a possibly singular multivariate normal serves as an example of tneir use. Applications are presented for a system of seemingly unrelated regressions and for the Linear Expenditure System. It is shown that commonly used iterative estimation schemes for these models coincide with a generalized method of scoring in the maximum likelihood framework.

Revision November 1982

THE USE OF GENERALIZED INVERSES IN RESTRICTED MAXIMUM LIKELIHOOD[1]

1. Introduction

This paper recalls and derives results in matrix calculus and shows their usefulness in maximum likelihood estimation. The calculus results are in the field of Moore-Penrose inverses and Kronecker products (section 2), and matrix differentation (section 3).
The power of Moore-Penrose inverses in restricted maximum likelihood estimation is demonstrated in section 4.
We then give some applications to well-known ML estimation problems.
As a first example, section 5 shows how the symmetry restriction on the covariance matrix is easily incorporated in ML estimation of a multivariate normal distribution without using "elimination" or "duplication" matrices. The treatment of a singular multivariate normal follows the same lines and produces an elegant generalization.
For a (possibly nonlinear) Zellner-type seemingly unrelated regressions model, we readily obtain the asymptotic covariance matrix of the ML estimator and show that the usual numerical method of iterated (nonlinear) GLS coincides with the method of scoring (section 6).
Allowing for a singular covariance structure, we can in the same framework study the estimation of allocation models without the need to drop one equation from the system.

---

[1] Earlier versions of this paper were presented at the 1982 Econometric Society Winter Symposium in Chiasso and at the 1982 Econometric Society European Meeting in Dublin. The comments of H. Neudecker and of the editor and referee of KM are gratefully acknowledged. Any remaining errors are mine. I am indebted to Mischa Vinkestijn for the conscientious typing of the manuscript.

As an example, in section 7 we analyze the nonlinear estimation problem that arises in the Linear Expenditure System, and discuss some methods proposed for its solution.

2. Moore-Penrose inverses and Kronecker products[2]

The *vec*-operator transforms an arbitrary matrix into a column vector by stacking its columns. For conformable matrices A,B we have

$$\text{trace } A'B = (\text{vec } A)'\text{vec } B \tag{2.1}$$

The *Kronecker product* of (m,n)-matrix $A = (a_{ij})$ and (p,q)-matrix B is the (mp,nq)-matrix $A \otimes B = (a_{ij}B)$.

The important relation between vecs and Kronecker products is

$$\text{vec } ABC = (C' \otimes A) \text{ vec } B \tag{2.2}$$

The *Schur product* of (m,n)-matrix $A = (a_{ij})$ and (m,n)-matrix $B = (b_{ij})$ is the (m,n)-matrix $A * B = (a_{ij} b_{ij})$.

The Schur product is commutative and for conformable a,b,c,d,E we have

$$c'(a*b) = a'(b*c) = b'(c*a) \tag{2.3}$$

and

$$(a*b)'E(c*d) = a'(bc'*E)d \tag{2.4}$$

---

[2] Throughout this paper, lower case characters denote column vectors or scalars and upper case characters denote matrices; primes indicate transposition.

The *Khatri-Rao product* (Khatri and Rao (1968)) of (m,n)-matrix $A = (a_1 \dots a_n)$ and (p,n)-matrix $B = (b_1 \dots b_n)$ is the (mp,n)-matrix $A \odot B = (a_1 \otimes b_1 \dots a_n \otimes b_n)$.
It is related to the Schur product by

$$(a'B) * c' = a'(B \odot c') \tag{2.5}$$

The *commutation matrix* K_{mn} is the unique (mn,mn)-matrix which for arbitrary (m,n)-matrix A satisfies

$$K_{mn} \text{ vec } A = \text{vec } A' \tag{2.6}$$

This matrix is extensively studied in Magnus and Neudecker (1979).
Its name is inspired by its effect on conformable Kronecker products:

$$K_{mn}(A \otimes B)K_{st} = B \otimes A \tag{2.7}$$

for (n,s)-matrix A and (m,t)-matrix B.
The (n^2,n^2)-matrix K_{nn} is also written K_n. It is symmetric and orthogonal, which implies that

$$\tfrac{1}{2}(I_{n^2} - K_n) \text{ and } \tfrac{1}{2}(I_{n^2} + K_n) \text{ are symmetric and idempotent} \tag{2.8}$$

Furthermore, for (n,n)-matrix A we have

$$K_n(A \otimes A) = (A \otimes A)K_n \tag{2.9}$$

For arbitrary (m,n)-matrix A, any (n,m)-matrix X that satisfies AXA = A is a *generalized inverse* of A and denoted A^-. The *Moore-Penrose inverse* of A is the unique (n,m)-matrix A^+ that satisfies

$$\text{(i) } AA^+A = A \qquad \text{(iii) } A^+A \text{ is symmetric} \qquad \text{(ii) } A^+AA^+ = A^+ \qquad \text{(iv) } AA^+ \text{ is symmetric} \tag{2.10}$$

The properties of A^- and A^+ and related matrices have been studied by a great number of authors; we recall the following properties of the Moore-Penrose inverse (e.g. Boullion and Odell (1971) or Rao and Mitra (1971)):

$A^+A = I$ iff A has full column rank (2.11)

$A^+ = (A'A)^+A' = A'(AA')^+$ (2.12)

$(AA')^+AA' = AA^+$ and $A'A(A'A)^+ = A^+A$ (2.13)

$(A')^+ = (A^+)'$ and $(A^+)^+ = A$ (2.14)

rank A = rank A^+ = rank A^+A = trace A^+A (2.15)

AA^+ projects on the column space of A (2.16)

$I - A^+A$ projects on the kernel of A (2.17)

AA^+, A^+A, $I-AA^+$ and $I-A^+A$ are symmetric and idempotent (2.18)

$(GAH)^+ = H(GAH)^+G$ if G and H are symmetric and idempotent (2.19)

the Euclidean distance $\|Ax - b\|$ reaches its minimal value for $x = A^+b + (I-A^+A)z$, z an arbitrary vector; the system Ax=b is consistent iff $AA^+b = b$ (2.20)

As a new result, we prove

<u>THEOREM 2.1</u>: For any (n,n)-matrix A we have

$$I_{n^2} - \begin{pmatrix} I_{n^2} - K_n \\ A' \otimes I_n \end{pmatrix}^+ \begin{pmatrix} I_{n^2} - K_n \\ A' \otimes I_n \end{pmatrix} =$$

$$= \tfrac{1}{2}(I_{n^2} + K_n) \left[(I_n - AA^+) \otimes (I_n - AA^+) \right] \tag{2.21}$$

Proof: Omitting the indices on I and K, we define

$$R = \begin{pmatrix} I - K \\ A' \otimes I \end{pmatrix} \text{ and } P = \tfrac{1}{2}(I + K) \left[(I - AA^+) \otimes (I - AA^+)\right].$$

From (2.9), we have $P = \left[(I-AA^+) \otimes (I - AA^+)\right]\tfrac{1}{2}(I + K)$,

and with this result one verifies that

(a) $RP = 0$ and

(b) P is idempotent and symmetric.

In view of (2.17), this proves the theorem if

(c) $Rx = 0$ implies $Px = x$.

So let $Rx = 0$, and X be the (n,n)-matrix such that $x = \text{vec } X$.

We obtain $X = X'$ and $XA = 0$, to find

$$Px = P\text{vec } X = \tfrac{1}{2}(I + K)\text{vec} \left\{(I - AA^+)X(I - AA^+)\right\}$$

$$= \tfrac{1}{2}(I + K) \text{ vec}X = x$$

which proves (c). □

COROLLARY: All symmetric solutions X to the matrix equation XA = 0 are given by

$$\text{vec } X = \tfrac{1}{2}(I + K)\left[(I - AA^{+}) \otimes (I - AA^{+})\right] z \qquad (2.22)$$

with z an arbitrary n^2-vector.

3. Matrix differentials and derivatives

A collection of appropriate propositions on matrix derivatives is provided in Dhrymes (1978).

Mainly for notational convenience, we restate some important results using differentials in stead of derivatives. For a continuously differentiable scalar function of a vector argument, the two are related by

$$\frac{\partial f}{\partial x} = a' \qquad \text{iff} \quad df = a'dx \qquad \text{for all } dx \qquad (3.1)$$

$$\frac{\partial^2 f}{\partial x \partial x'} = \tfrac{1}{2}(A + A') \quad \text{iff} \quad d^2 f = (dx)'A(dx) \quad \text{for all } dx. \qquad (3.2)$$

For details on matrix differentials and related results, see Neudecker (1967, 1969).

If X and Y are conformable continuously differentiable matrix functions, we have

$$d(XY) = (dX)Y + X(dY) \qquad (3.3)$$

$$d \text{ trace } X = \text{trace } dX \qquad (3.4)$$

$$d \text{ vec } X = \text{vec } dX \qquad (3.5)$$

and at points where X is nonsingular

$$d \ln |X| = \text{trace } X^{-1} dX \tag{3.6}$$

$$d X^{-1} = -X^{-1} (dX) X^{-1} \tag{3.7}$$

For the special matrix products $\otimes$, $*$ and $\odot$ the analogon of (3.3) holds.

A generalization of (3.7) to (m,n)-matrices X of locally constant rank[3] was given by Golub and Pereyra (1973):

$$dX^{+} = - X^{+}(dX)X^{+} + (I - X^{+}X)(dX')(XX')^{+} + (X'X)^{+}(dX')(I - XX^{+}) \tag{3.8}$$

A generalization of (3.6) to arbitrary (m,n)-matrices X is given in Theorem 3.1, which as far as we know is new. It is based on the *singular value decomposition* (e.g. Rao (1973, p. 42)):

an arbitrary (m,n)-matrix A of rank r can always be written $A = U\Lambda V'$ where Λ is a diagonal (r,r)-matrix with positive diagonal elements, and (m,r)-matrix U and (n,r)-matrix V satisfy $U'U = V'V = I_r$.

One verifies that the diagonal elements of Λ (i.e. the *singular values* λ_i) equal the square roots of the r positive eigenvalues of AA' (or A'A), while the columns of U are the corresponding orthonormal eigenvectors of AA' and the columns of V those of A'A.

The Moore-Penrose inverse of $A = U\Lambda V'$ is $A^{+} = V\Lambda^{-1}U'$, and we have

$$AA^{+} = UU' \quad \text{and} \quad A^{+}A = VV'.$$

---

[3] The matrix function $X(\alpha)$ has locally constant rank in α_0 if rank $X(\alpha)$ = rank $X(\alpha_0)$ for all α in some open neighbourhood of α_0.

THEOREM 3.1: For any continuously differentiable (m,n)-matrix function X, we have at points where rank X = r:

$$d \log \prod_{i=1}^{r} \lambda_i(X) = \text{trace } X^+ dX, \tag{3.9}$$

where $\lambda_i(X)$ denote the r largest singular values of X.

REMARK: In the sequel we will only need (3.9) for square symmetric nonnegative definite matrices of constant rank, having a fixed kernel. For such matrices the proof of (3.9) can be derived in a straightforward manner from (3.6). However, the more general Theorem 3.1 may be of use in other applications.

Proof (cf. Neudecker (1967)): Consider a point where rank X = r and all r nonzero eigenvalues of X'X are simple roots of its characteristic polynomial. The coefficients of this polynomial are continuously differentiable functions of the elements of X, and thus its simple roots and the associated eigenvectors of X'X are continuously differentiable functions in view of the Implicit Function Theorem. The same holds for eigenvectors of XX', and thus the matrices Λ, U and V of the singular value decomposition $X = U\Lambda V'$ are continuously differentiable at points where X'X has r simple nonzero roots.

Moreover, in such points we have

$$d\Lambda = d(U'XV) = (dU)'XV + U'(dX)V + U'X(dV) \tag{3.10}$$

$$= (dU)'U\Lambda + U'(dX)V + \Lambda V'(dV).$$

The identity $U'U = I_r$ implies that the diagonal elements of (dU)'U vanish, so $(dU)'U\Lambda = 0$. Similarly, $\Lambda V'(dV) = 0$, and we find

$$d \log \prod_{i=1}^{r} \lambda_i(X) = d \log |\Lambda| = \text{trace}\, \Lambda^{-1}(d\Lambda)$$

$$= \text{trace}\, \Lambda^{-1} U'(dX)V = \text{trace}\, X^{+}(dX). \qquad (3.11)$$

Now consider the possibility of multiple roots.
Because the coefficients of the characteristic polynomial of $X'X$ are continuously differentiable functions, so are the following expressions in its roots $\lambda_1^2 \ldots \lambda_n^2$:

$\lambda_1^2 + \lambda_2^2 + \ldots + \lambda_n^2$; $\lambda_1^2\lambda_2^2 + \lambda_1^2\lambda_3^2 + \ldots + \lambda_{n-1}^2\lambda_n^2$; ... ;

$\lambda_1^2\lambda_2^2 \ldots \lambda_n^2$. The identification of the r largest roots is locally unambiguous. The stated expressions are continuously differentiable functions of the corresponding expressions in the n-r smallest and r largest roots respectively, and yield simple solutions in the latter.

Thus, irrespective of multiple roots, the product of the r largest roots is a continuously differentiable function in points where rank $X = r$. So by a continuity argument, (3.9) also holds at points where multiple roots occur. □

4. Restricted maximum likelihood estimation

Maximum Likelihood Estimation subject to equality restrictions was rigorously studied by Aitchison and Silvey (1958). They stated conditions both on the loglikelihood $L_T(\theta)$ and the restrictions $g(\theta) = 0$ for the existence of a consistent restricted ML estimator. They also discussed its asymptotic distribution and an iterative method to solve the first order equations.

Let $B_T = - E \frac{\partial^2 L_T}{\partial\theta\partial\theta'}$ denote the information matrix, and define $B(\theta) = \lim_{T\to\infty} \frac{1}{T} B_T(\theta)$. The matrix of first order derivatives of the restrictions is denoted $G(\theta) = \frac{\partial}{\partial\theta} g(\theta)$. The true parameter vector θ_0 is known to satisfy $g(\theta_0) = 0$.

The well-known result is that under suitable assumptions the vector $\hat{\theta}$ that solves the first order conditions[4]

$$\frac{\partial L_T}{\partial \theta}(\hat{\theta}) + \lambda' G(\hat{\theta}) = 0 \quad \text{and} \quad g(\hat{\theta}) = 0 \tag{4.1}$$

is consistent and asymptotically normal with asymptotic covariance matrix equal to the NW submatrix of the inverted bordered asymptotic information matrix $\begin{pmatrix} B & G' \\ G & 0 \end{pmatrix}^{-1}$ evaluated at θ_0.

Here we make all (regularity) assumptions given in Aitchison and Silvey (1958), except for two rank conditions they impose:

1. $B(\theta_0)$ is positive definite
2. $G(\theta_0)$ has full row rank.

The first assumption was relaxed by Rothenberg (1973, p.22) who showed that we only need[5]

1^*. $B(\theta_0) + G'(\theta_0)G(\theta_0)$ is positive definite

This section will provide an elegant expression for the asymptotic covariance matrix of $\hat{\theta}$, and slightly relax assumption 2 to the assumption

2^*. $G(\theta_0)$ has locally constant rank.

---

[4] λ denotes a vector of Lagrangian multipliers.

[5] We need 1^* for identification of the full parameter vector; if it is not satisfied, one may proceed as in Rao and Mitra (1971, p. 201-203) to obtain results for the estimable functions.

THEOREM 4.1: The first order conditions (4.1) on the MLE $\hat{\theta}$ are equivalent with

$$\frac{\partial L}{\partial \theta}(\hat{\theta}).[I - G^{+}(\hat{\theta})G(\hat{\theta})] = 0 \quad \text{and } g(\hat{\theta}) = 0 \tag{4.2}$$

and under assumptions 1^{*},2, the asymptotic covariance matrix of $\hat{\theta}$ equals

$$[(I - G^{+}G)B(I - G^{+}G)]^{+} \quad \text{evaluated at } \theta_0 \tag{4.3}$$

Proof: The first equation of (4.1) admits of a solution for the vector λ of Lagrangian multipliers iff the first equation of (4.2) is satisfied by $\hat{\theta}$ (cf. property (2.20) above).

Rothenberg (1973) showed that on the assumptions 1^{*},2, the asymptotic covariance matrix of $\hat{\theta}$ equals

$$(B + G'G)^{-1} - (B + G'G)^{-1}G'[G(B + G'G)^{-1}G']^{-1}G(B + G'G)^{-1}$$

evaluated at θ_0. One verifies that this expression equals the one stated above by checking the defining properties of the Moore-Penrose inverse. □

THEOREM 4.2: Assumption 2 may be replaced by 2^{*} without affecting the results of Theorem 4.1.

Proof: The derivation of (4.2) does not need assumption 2. Now suppose rank $G(\theta_0) = r$. Then $G(\theta_0)$ contains r independent rows, which may be taken to be the first r rows. As $G(\theta)$ is continuous (one of the assumptions in Aitchison and Silvey (1958)), these rows $G_r(\theta)$ are independent in some neighbourhood U_1 of θ_0. As $G(\theta)$ is of locally constant rank in θ_0, we have rank $G(\theta) = r$ in some neighbourhood U_2 of θ_0. Thus in $U_1 \cap U_2$ the first r restrictions $g_r(\theta) = 0$ are equivalent to the full set $g(\theta) = 0$. Asymptotically, $\hat{\theta}$ will fall in $U_1 \cap U_2$ with unit probability. Thus its asymptotic covariance matrix equals $[(I - G_r^+ G_r) B (I - G_r^+ G_r)]^+$ evaluated at θ_0 if $B + G_r' G_r$ is positive definite. Because $G^+ G = G_r^+ G_r$ and G has the same kernel as G_r, Theorem 4.1 is not affected if assumption 2 is replaced by 2^*. □

If the restrictions g are linear (i.e. if $G(\theta)$ is constant) we readily obtain an iterative scheme to find a numerical solution to the first order conditions (4.2). A generalization of the *method of scoring* (e.g. Rao (1973, p. 367)) yields:

1) choose a starting value θ_1 that satisfies $g(\theta_1) = 0$
2) calculate θ_2, θ_3, ... through

$$\theta_{k+1} = \theta_k + [(I - G^+G) B_T (I - G^+G)]^+ \left(\frac{\partial L_T}{\partial \theta}\right)' \qquad k = 1,2, \ldots \qquad (4.4)$$

where the RHS is evaluated in θ_k.

Note that the linearity of g is required to ensure that $g(\theta_k) = 0$ for $k = 2,3, \ldots$.

5. ML estimation of the multivariate normal

It has been stressed by various authors (e.g. Richard (1975), Balestra (1976), Magnus and Neudecker (1980)) that in estimating a covariance matrix one must take the symmetry restriction into account in order to obtain the correct asymptotic variance of the estimator. Consider a sample $y_1 \dots y_T$ of the n-variate normal $N(\mu, \Omega)$ with nonsingular Ω. The symmetry restriction on Ω can be stated as[6]

$$(I - K) \text{ vec } \Omega = 0 \tag{5.1}$$

In view of Theorem 4.2, we need not bother about the singularity of I - K.

Subject to this linear restriction, we maximize the loglikelhood

$$L = \text{constant} - \tfrac{1}{2}T \log|\Omega| \quad - \tfrac{1}{2} \sum_t (y_t-\mu)'\Omega^{-1}(y_t-\mu) \tag{5.2}$$

Taking differentials, we obtain

$$dL = -\tfrac{1}{2}\, T \text{ trace } [\Omega^{-1} - \Omega^{-1} \frac{1}{T} \sum_t (y_t-\mu)(y_t-\mu)'\Omega^{-1}](d\Omega)$$

$$+ \sum_t (y_t-\mu)'\Omega^{-1}(d\mu) \tag{5.3}$$

and

$$E\, d^2L = -\tfrac{1}{2}T \text{ trace } \Omega^{-1}(d\Omega)\Omega^{-1}(d\Omega) - T(d\mu)'\Omega^{-1}(d\mu) \tag{5.4}$$

---

[6] We omit indices on I and K if their dimensions are unambiguous.

implying

$$\frac{\partial L}{\partial \mu} = \sum_t (y_t - \mu)'\Omega^{-1} \tag{5.5}$$

$$\frac{\partial L}{\partial \mathrm{vec}\,\Omega} = -\tfrac{1}{2}T\left[\mathrm{vec}(\Omega^{-1} - \Omega^{-1}\frac{1}{T}\sum_t (y_t-\mu)(y_t-\mu)'\Omega^{-1})\right]' \tag{5.6}$$

$$B_T = \begin{pmatrix} T\,\Omega^{-1} & 0 \\ 0 & \tfrac{1}{2}T\,\Omega^{-1} \otimes \Omega^{-1} \end{pmatrix} \tag{5.7}$$

With $G = \begin{pmatrix} 0 \\ I-K \end{pmatrix}$, we have $(I-G^{+}G) = \begin{pmatrix} I & 0 \\ 0 & \tfrac{1}{2}(I+K) \end{pmatrix}$

because of (2.8). Substitution into (4.2) and (4.3) now yields the familiar ML solution $\hat{\mu} = \frac{1}{T}\sum_t y_t$ and $\hat{\Omega} = \frac{1}{T}\sum_t (y_t - \mu)(y_t - \hat{\mu})'$ with asymptotic covariance matrix of $\begin{pmatrix} \hat{\mu} \\ \mathrm{vec}\,\hat{\Omega} \end{pmatrix}$ given by

$$V = \begin{pmatrix} \Omega & 0 \\ 0 & (I+K)(\Omega \otimes \Omega) \end{pmatrix} \tag{5.8}$$

Use is made of the identity

$$\left[\tfrac{1}{2}(I+K)(\Omega^{-1} \otimes \Omega^{-1})\tfrac{1}{2}(I+K)\right]^{+} = \tfrac{1}{2}(I+K)(\Omega \otimes \Omega)\tfrac{1}{2}(I+K) = \tfrac{1}{2}(I+K)(\Omega \otimes \Omega).$$

The result conforms to the formulae obtained by other authors, but in deriving (5.8) we did not need devices like "elimination" or "duplication" matrices. The derivation clearly sets out the method adopted in this paper for more complicated problems in Sections 6 and 7.

The present approach can easily incorporate the case of a singular covariance matrix Ω. Consider the case that Ω looses m in rank, and the known kernel of Ω is spanned by the columns of some (n,m)-matrix A. Instead of removing the singularity by discarding m elements of the sample vectors y_t, one may wish to treat all elements of y_t symmetrically and impose the singularity of Ω in estimation.
The appropriate loglikelihood now reads[7] (Rao (1973, p. 528)):

$$L = \text{constant} - \tfrac{1}{2} T \log \prod_{i=1}^{n-m} \lambda_i(\tilde{\Omega}) - \tfrac{1}{2}\sum_t (y_t-\mu)'\tilde{\Omega}^+(y_t-\mu) \tag{5.9}$$

where $\lambda_i(\tilde{\Omega})$ denote the positive eigenvalues of $\tilde{\Omega} = (I-AA^+)\Omega(I-AA^+)$. As Ω is symmetric and nonnegative definite, the λ_i coincide with its singular values.
Expression (5.9) should now be maximized subject to the symmetry and singularity restrictions on Ω:

$$\begin{pmatrix} I - K \\ A' \otimes I \end{pmatrix} \text{vec}\, \Omega = 0 \tag{5.10}$$

Taking differentials, we obtain

$$dL = -\tfrac{1}{2} T \,\text{trace} \left[\tilde{\Omega}^+ - \tilde{\Omega}^+ \frac{1}{T} \sum_t (y_t-\mu)(y_t-\mu)'\tilde{\Omega}^+ \right] (d\Omega)$$

$$+ \sum_t (y_t-\mu)'\tilde{\Omega}^+(d\mu) \tag{5.11}$$

---

[7] $\tilde{\Omega}$ is used instead of Ω in order to guarantee that L is continuously differentiable with respect to μ, vec Ω. We have $d\tilde{\Omega}^+ = -\tilde{\Omega}^+(d\Omega)\tilde{\Omega}^+$.

and

$$E\, d^2L = -\tfrac{1}{2} T \operatorname{trace} \tilde{\Omega}^+ (d\Omega) \tilde{\Omega}^+ (d\Omega) - T(d\mu)' \tilde{\Omega}^+ (d\mu) \tag{5.12}$$

Theorem 2.1 readily gives $I-G^+G$ for this example, so we can proceed to substitute our results into (4.2) and (4.3) to obtain the ML solution $\hat{\mu} = \frac{1}{T}\sum_t y_t$ and $\hat{\Omega} = \frac{1}{T}\sum_t (y_t-\hat{\mu})(y_t-\hat{\mu})'$ with asymptotic covariance matrix of $\begin{pmatrix} \hat{\mu} \\ \operatorname{vec} \hat{\Omega} \end{pmatrix}$ again given by (5.8), because $\tilde{\Omega} = \Omega$ at the true parameter values.

6. Seemingly unrelated regressions

Consider a possibly nonlinear seemingly unrelated regressions model

$$y_t = f_t(\beta) + u_t \tag{6.1}$$

$$u_t \text{ i.i.d. and } N(0,\Psi) \tag{6.2}$$

where Ψ is a nonsingular (n,n)-matrix of contemporaneous covariances.

We readily obtain the loglikelihood

$$L = \text{constant} - \tfrac{1}{2} T \log |\Psi| - \tfrac{1}{2} \sum_t (y_t - f_t(\beta))' \Psi^{-1} (y_t - f_t(\beta)) \tag{6.3}$$

Assume that f_t is twice continuously differentiable with $\frac{\partial f_t}{\partial \beta} = X_t$, and define $e_t = y_t - f_t(\beta)$. Taking differentials, we obtain

$$dL = -\tfrac{1}{2} T \operatorname{trace}\left[\Psi^{-1} - \Psi^{-1} \frac{1}{T} \sum_t e_t e_t' \Psi^{-1}\right](d\Psi)$$

$$+ \sum_t e_t' \Psi^{-1} X_t (d\beta) \qquad (6.4)$$

and

$$E\, d^2L = -\tfrac{1}{2} T \operatorname{trace} \Psi^{-1}(d\Psi)\Psi^{-1}(d\Psi) - (d\beta)' \sum_t X_t' \Psi^{-1} X_t (d\beta) \qquad (6.5)$$

If the only restriction on the parameters is the symmetry of Ψ, the asymptotic covariance matrix for $\begin{pmatrix} \hat{\beta} \\ \operatorname{vec} \hat{\Psi} \end{pmatrix}$ is

$$\begin{pmatrix} \left(\frac{1}{T} \sum_t X_t' \Psi^{-1} X_t\right)^{-1} & 0 \\ 0 & (I+K)(\Psi \otimes \Psi) \end{pmatrix} \qquad (6.6)$$

which conforms to the result obtained by Magnus (1978) for the linear case.

Applying the generalized method of scoring as developed in Section 4, we obtain the recursions

$$\beta_{k+1} = \left(\sum_t X_t' \Psi_k^{-1} X_t\right)^{-1} \sum_t X_t' \Psi_k^{-1} y_t \qquad (6.7)$$

$$\Psi_{k+1} = \frac{1}{T} \sum_t e_t e_t' \quad \text{evaluated in } \beta_k \qquad (6.8)$$

These recursions describe[8] the commonly used procedure known as iterated (nonlinear) GLS.

7. Estimation of the Linear Expenditure System

A seemingly unrelated regressions system with a singular covariance matrix Ψ arises in allocation models, where the disturbances are known to satisfy a linear (adding-up) restriction. The common estimation procedure for such models is to drop one equation from the system and thus discard the adding-up restrictions. It is well-known (Barten (1969), Deaton (1975)) that in Maximum Likelihood Estimation of general nonlinear allocation models the choice of the equation to be dropped does not affect the results.

With the approach developed in section 5 we are able to derive formulae for the Maximum Likelihood problem without dropping any of the equations from the system. In most cases, adding up will impose more restrictions on the parameters than symmetry and singularity of the covariance matrix. As an example, we consider in this section the estimation of a well-known non-linear allocation model, the Linear Expenditure System (LES).

---

8 In fact (6.7), (6.8) describe two independent iterative schemes, viz. the sequences $(\Psi_0, \beta_1, \Psi_2, \beta_3, \ldots)$ and $(\beta_0, \Psi_1, \beta_2, \Psi_3, \ldots)$. The starting values usually satisfy $\Psi_0 = I$ and $\beta_0 = \beta_1$, implying that the two schemes are numerically the same.

The Maximum Likelihood estimation of the LES was studied e.g. by Parks (1971), Deaton (1975) and Ham (1978). The system[9] reads, in obvious notation,

$$y_{it} = \gamma_i p_{it} + \beta_i (m_t - \sum_{j=1}^{n} \gamma_j p_{jt}) + u_{it} \tag{7.1}$$

where $i = 1, \ldots, n$ indexes commodities and $t = 1, \ldots, T$ time.

We have $\sum_{i=1}^{n} y_{it} \equiv m_t$, $\sum_{i=1}^{n} \beta_i = 1$ and $\sum_{i=1}^{n} u_{it} \equiv 0$.

Collect the y_{it}, p_{it} and u_{it} in n-vectors y_t, p_t and u_t and define the parameter vectors $b = (b_1 \ldots b_n)$ and $c = (c_1 \ldots c_n)$. Using the Schur product $*$, the system (7.1) can now be written

$$y_t = (I - bs')c*p_t + bm_t + u_t \tag{7.2}$$

or equivalently

$$u_t = (I - bs')(y_t - c*p_t) \tag{7.3}$$

On the usual assumption that the u_t are i.i.d. and $N(0, \Psi)$ with Ψ a covariance matrix of rank n-1 that satisfies $s'\Psi = 0$, and writing $\breve{\Psi} = (I-ss^+)\Psi(I-ss^+)$, we have the loglikelihood

---

9 Deaton (1975) incorporates a trend in the β_i; for the sake of simplicity it is left out here.

$$L = \text{constant} \;\; -\tfrac{1}{2}\, T \log \prod_{i=1}^{n-1} \lambda_i(\tilde{\Psi})$$

$$-\tfrac{1}{2} \sum_t (y_t - c*p_t)'(I-sb')\tilde{\Psi}^+(I-bs')(y_t - c*p_t) \qquad (7.4)$$

Taking differentials, we obtain

$$dL = -\tfrac{1}{2}\, T \,\text{trace}\left[\tilde{\Psi}^+ - \tilde{\Psi}^+ \frac{1}{T} \sum_t e_t e_t' \;\Psi^+ \;\right](d\Psi)$$

$$+ \sum_t \left\{(y_t - c*p_t)'(I-sb')\tilde{\Psi}^+(I-bs')\right\} * p_t' \;(dc)$$

$$+ \sum_t (y_t - c*p_t)'(I-sb')\tilde{\Psi}^+(db)s'(y_t - c*p_t) \qquad (7.5)$$

and

$$E \;\; d^2L = -\tfrac{1}{2}\, T \,\text{trace}\, \tilde{\Psi}^+(d\Psi)\tilde{\Psi}^+(d\Psi)$$

$$- \sum_t (dc)' \left[p_t p_t' * \left\{(I-sb')\tilde{\Psi}^+(I-bs')\right\}\right] (dc)$$

$$- 2 \sum_t (y_t - c*p_t)'s \;(db)'\left\{\tilde{\Psi}^+(I-bs')\right\} \odot p_t' \;(dc)$$

$$- \sum_t (y_t - c*p_t)'s(db)'\tilde{\Psi}^+(db)s'(y_t - c*p_t) \qquad (7.6)$$

The restrictions on the parameters read

$$\begin{pmatrix} s' & 0 & 0 \\ 0 & 0 & s' \otimes I \\ 0 & 0 & I-K \end{pmatrix} \begin{pmatrix} b \\ c \\ \text{vec } \Psi \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{7.7}$$

and thus in this case we have

$$I-G^{+}G = \begin{pmatrix} I-ss^{+} & 0 & 0 \\ 0 & I_n & \\ 0 & 0 & \tfrac{1}{2}(I+K)\left[(I-ss^{+}) \otimes (I-ss^{+})\right] \end{pmatrix} \tag{7.8}$$

Upon substitution into (4.3), we obtain the asymptotic covariance matrix of the ML estimator $\begin{pmatrix} \hat{b} \\ \hat{c} \\ \text{vec } \hat{\Psi} \end{pmatrix}$,

$$V = \begin{pmatrix} H^{+} & 0 \\ 0 & (I+K)(\Psi \otimes \Psi) \end{pmatrix} \tag{7.9}$$

with $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}' & H_{22} \end{pmatrix}$ the square matrix of order 2n

formed by the (n,n)-blocks

$$H_{11} = \frac{1}{T}\sum_t \{s'(y_t - c*p_t)\}^2 \Psi^+$$

$$H_{12} = \frac{1}{T}\sum_t s'(y_t - c*p_t) \{\Psi^+(I-bs')\} \odot p_t' \qquad (7.10)$$

$$H_{22} = \frac{1}{T}\sum_t p_t p_t' * \{(I-sb')\Psi^+(I-bs')\}$$

Again, we may apply the generalized method of scoring to solve the ML estimation problem numerically. The recursion formula (4.4) specializes to

$$\begin{pmatrix} \hat{b}_{k+1} \\ \hat{c}_{k+1} \end{pmatrix} = H_k^+ \; \frac{1}{T}\sum_t \begin{pmatrix} s'(y_t - c_k*p_t)\Psi_k^+(y_t - c_k*p_t) \\ p_t*(I-sb_k')\Psi_k^+(I-b_k s')y_t \end{pmatrix} \qquad (7.11)$$

and

$$\Psi_{k+1} = \frac{1}{T}\sum_t (I-bs')(y_t - c_k*p_t)(y_t - c_k*p_t)'(I-sb') \qquad (7.12)$$

If, in (7.11), the off-diagonal blocks of H_k are discarded, we obtain GLS-type recursions for $\hat{b}$ and $\hat{c}$:

$$\hat{b}_{k+1} = [\sum_t \{s'(y_t - c_k*p_t)\}^2]^{-1} \sum_t s'(y_t - c_k*p_t)(y_t - c_k*p_t) \qquad (7.13)$$

$$\hat{c}_{k+1} = [\sum_t p_t p_t' * (I-sb_k')\Psi_k^+(I-b_k s')]^{-1} \sum_t p_t * (I-sb_k')\Psi_k^+(I-b_k s')y_t$$

(7.14)

These describe the original estimation procedure proposed by Stone (Parks (1971), Deaton (1975)). The alternative suggested by Parks (1971) uses the full[10] matrix H_k in (7.11).
Deaton (1975) developed a "ridge-walking algorithm", exploiting the fact that if c is fixed, the first order conditions are linear in b. This provides the possibility to concentrate b out of the loglikelihood by solving a linear system (yielding 7.13), and next apply the method of scoring to the concentrated loglikelihood. The resulting iterative scheme needs inversions of order n only.
Note that concentrating out Ψ does not yield any savings in the order of inversions, because the off-diagonal blocks in (7.9) are zero and its partitioned inverse led to (7.12).
Finally, Ham (1978) suggests to maximize with numerical optimization routines (i.e. without analytically evaluated derivatives) the twice concentrated loglikelihood obtained after concentrating out Ψ and b. If the dimensionality of the system is high, this might prove efficient in terms of computer time. However, in a recent study of Belsley (1980) the advantage of analytically computed derivatives in numerical optimization was found to be considerable.
The nonlinear allocation model studied in Don (1982) is very similar to the LES. The numerical results reported there were obtained from an analogon of (7.11), (7.12). Some computer time was saved by using the same H_k^+ -matrix in steps k, k+1 and k+2 for k = 6,9,12, ...

---

[10] That is, a similar but nonsingular matrix of order 2n-2.

8. Concluding remarks

We have shown how the calculus of generalized inverses and matrix differentials can be usefully applied to restricted maximum likelihood problems. The usual formulae need slight modifications to handle the cases of locally superfluous restrictions and singular parameter covariance matrix with known kernel.

It is not claimed that any additional information can be drawn from these exercises in algebra, but I feel that there is some merit in treating all variables, parameters and restrictions in a symmetrical way.

We found that several commonly used iterative schemes for well-known ML estimation problems actually coincide with a generalized method of scoring . This should not be understood as an unconditional recommendation for the use of scoring methods. In applying such iterative schemes, it is important to check whether the likelihood really does increase at each step of the iteration. If it does not, a change of step size direction or both is in order. Applying a linear search along the favoured direction in each step can be overdoing it: in most cases, the optimal step size is very close to unity. Changes of direction are a succesful feature of the Marquardt method; see e.g. Bard (1974) for a discussion and further references.

REFERENCES

Aitchison, J. and S.D. Silvey (1958). Maximum-likelihood estimation of parameters subject to restraints, Annals of Mathematical Statistics 29, 813-828.

Balestra, P. (1976). La Dérivation Matricielle. Collection de l'IME, 12, Paris, Sirey.

Bard, Y. (1974). Nonlinear Parameter Estimation. New York, Academic Press.

Barten, A.P. (1969). Maximum likelihood estimation of a complete system of demand questions, European Economic Review 1, 7-73.

Belsley, D.A. (1980). On the efficient computation of the nonlinear full-information maximum-likelihood estimator, Journal of Econometrics 14, 203-225.

Bouillon, T.L. and P.L. Odell (1971). Generalized Inverse Matrices. New York, Wiley.

Dhrymes, P.J. (1978). Mathematics for Econometrics. New York, Springer.

Don, F.J.H. (1982). Restrictions on variables, Journal of Econometrics 18, 369-393.

Golub, G.H. and V. Pereyra (1973). The differentiation of pseudo-inverses and nonlinar least squares problmes whose variables separate, SIAM Journal of Numerical Analysis 10, 413-432.

Ham, J.C. (1978). A note on the efficient estimation of the Linear Expenditure System, Journal of the American Statistical Association 73, 208-210.

Khatri, C.G. and C.R. Rao (1968). Solutions to some functional equations and their applications to characterization of probability distributions, Sankhyā A 30, 167-180.

Magnus, J.R. (1978). Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix, Journal of Econometrics 7, 281-312.

Magnus, J.R. and H. Neudecker (1979). The commutation matrix: some properties and applications, Annals of Statistics 7, 381-394.

Magnus, J.R. and H. Neudecker (1980). The elimination matrix: some lemmas and applications. SIAM Journal of Algebraic and Dicrete Methods 1, 422-449.

Neudecker, H. (1967). On matrix procedures for optimizing differentiable scalar functions of matrices, Statistica Neerlandica 21, 101-107.

Neudecker, H. (1969). Some theorems on matrix differentiation with special reference to Kronecker matrix products, Journal of the American Statistical Association 64, 953-963.

Parks, R.W. (1971). Maximum likelihood estimation of the Linear Expenditure System, Journal of the American Statistical Association 66, 900-903.

Rao, C.R. (1973). Linear Statistical Inference and Its Applications, 2nd ed. New York, Wiley.

Rao, C.R. and S.K. Mitra (1971). Generalized Inverse of Matrices and its Applications. New York, Wiley.

Richard, J.F. (1975). A note on the information matrix of the multivariate normal distribution, Journal of Econometrics 3, 57-60.

Rothenberg, Th. J. (1973). Efficient Estimation with A Priori Information, New Haven, Yale University Press.