DISCUSSION PAPER NO. 96

THE GENERALIZED SYMMETRIC EIGENPROBLEM IN MULTIVARIATE STATISTICAL MODELS

Claude Cohen

Revised June 1974

Graduate School of Management

Northwestern University

Evanston, Illinois 60201

Abstract

The solution of the generalized symmetric eigenproblem $Ax = \lambda Bx$ is required in many multivariate statistical models, viz. canonical correlation, discriminant analysis, multivariate linear model, limited information maximum likelihood. The problem can be solved by two efficient numerical algorithms: Cholesky decomposition or singular value decomposition. Practical considerations for implementation are also discussed.

Key words

Generalized symmetric eigenproblem; canonical analysis; discriminant analysis; multivariate linear model; LIML; Cholesky; singular value.

1. INTRODUCTION AND PRELIMINARIES

The increasing practical use of multivariate statistical models is linked with the availability of many statistical programming systems. For a survey, see [17]. At times, however, and with no forewarning, computer programs which implement estimation procedures will produce unreasonable results from reasonable data. Only recently has an effort been made to graft numerical algorithms of linear algebra to statistical procedures. The literature concerning numerical linear algebra has vastly grown. For a perspective on the state of the art, see [20]. Statisticians are just beginning to recognize in the numerical analyst's world the algorithms they need. The least-squares/linear regression problem has been thoroughly investigated by several authors and an evaluation of some of the problems encountered can be found in [19]. The purpose of this paper is to approach the models of canonical correlation, discriminant analysis, multivariate linear hypothesis, and limited information maximum likelihood as extremum eigenvalue problems and make use of computationally efficient algorithms familiar to numerical analysts.

<u>1.1</u> If X is a $n \times p$ matrix, $n \geq p$, rank $(X) = p$ and a vector y are given, the least-squares/linear regression problem

$$(1) \qquad \min_{b} \| Xb - y \|$$

(where $\| \ldots \|$ denotes Euclidean norm) can be solved by <u>Householder's transformation</u> to obtain a decomposition

$$(2) \qquad X = QR$$

where Q is orthogonal, i.e. $Q'Q = I$ and

(3) $$R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}, \quad \tilde{R} \text{ upper triangular.}$$

An efficient algorithm for solving (1) is given in [2] where Q is obtained as the product of p Householder transformations. A square matrix of the form $H = I - 2ww'$, where $w'w = 1$, is said to be a Householder transformation. It is easy to check that H is a symmetric, orthogonal transformation. The problem $\min_b ||Xb - y||$ is transformed to $\min_b ||Q'Xb - Q'y||$ which reduces to solving a triangular system of equations.

(4) $$\tilde{R}b = z \quad , \quad z = Q'y$$

In statistical work, the solution of (1) is computed as $b = (X'X)^{-1}X'Y$, where the computation of the cross-product matrix and its inverse may lead to serious numerical difficulties. The best way to compute $(X'X)^{-1}$ is by <u>Cholesky decomposition</u> (CD) such that $(X'X) = LL'$, L lower triangular. See e.g. [20]. Denoting the elements of $(X'X)$ by a_{ij}, we have

$$\ell_{11} = \sqrt{a_{11}} \quad , \quad \ell_{i1} = a_{i1}/\ell_{11} \qquad i = 2,\dots,p$$

then, for $i = 2,\dots,p$

(5) $$\begin{cases} \ell_{ii} = \sqrt{a_{ii} - \sum_{j=1}^{i-1} \ell_{ij}^2} \\ \\ \ell_{ij} = (a_{ij} - \sum_{k=1}^{j-1} \ell_{ik}\ell_{jk})/\ell_{ii} \qquad j = i+1,\dots,p \end{cases}$$

Note that Householder's transformation of X yields $L = \tilde{R}'$ since

$$\tilde{R}'\tilde{R} = (Q'X)'Q'X = X'X$$

1.2 It is well known that, if A and B are symmetric matrices

(6) $$\min_{||x||=1} \frac{x'Ax}{x'Bx}$$

corresponds to the smallest eigenvalue of $(A - \lambda B)x = 0$. If B is positive definite the solution of the generalized symmetric eigenproblem (GSE)

(7) $$Ax = \lambda Bx$$

can be reduced to the standard symmetric eigenproblem

(8) $$Cy = \lambda y$$

where* $C = L^{-1}AL^{-t}$ and $y = L'x$ and $LL' = B$ is the Cholesky decomposition of B into upper/lower triangular matrices. Note that the eigenvalues of (7) and (8) are the same. An efficient algorithm for solving (7) can be found in [13]. In statistical work, the solution of (7) is often obtained from

(9) $$B^{-1}Ax = \lambda x$$

which is, in general, a non-symmetric eigenproblem requiring more computational work.

1.3 If M is any $p \times q$ real matrix, $p \leq q$, then there exist two orthogonal matrices S and T such that

(10) $$S'MT = (D,0)$$

---

*We use L^{-t} instead of the cumbersome $(L^{-1})'$.

where $\qquad D = \text{diag}\,(d_i) \qquad i = 1,\ldots,p$

$\qquad 0$ is the $p \times (q-p)$ zero matrix

and $\qquad d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$

are the <u>singular values of</u> M, i.e. the non-negative square roots of the eigenvalues of MM'. S and T are the eigenvector matrices of MM' and M'M respectively. Equivalently,

(11) $$M = S(D,0)T'$$

and if rank(M) = k, $d_{k+1} = \ldots = d_p = 0$. See [5,6].

An efficient algorithm to obtain the <u>singular value decomposition</u> (SVD) of M can be found in [11].

2. STATISTICAL APPLICATIONS

2.1 Canonical Correlation

Canonical analysis considers the relationship between two sets of variables $X_1, X_2, \ldots, X_p$ and $Y_1, Y_2, \ldots, Y_q$, $p \leq q$ and reduces the study of $p \times q$ correlations to the study of p canonical correlations $\lambda_1, \ldots, \lambda_p$ between two sets of canonical variables $U_1, U_2, \ldots, U_p$ and $V_1, V_2, \ldots, V_q$. If the X_i's and Y_j's are jointly distributed with sample variance-covariance matrix S partitioned as

$$S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}$$

then the variance-covariance matrix of $U_i = \alpha'_{(i)}X$ and $V_j = \beta'_{(j)}Y$, where $\alpha_{(i)}$ and $\beta_{(j)}$ are column vectors, can be written as

$$\text{(12)} \qquad \left(\begin{array}{c|c|c} I_p & \Lambda & 0 \\ \hline \Lambda & \multicolumn{2}{c}{I_q} \\ \cline{1-1} 0 & \multicolumn{2}{c}{} \end{array}\right) \qquad \Lambda = \mathrm{diag}(\lambda_i) \quad i=1,\ldots,p \ .$$

The sought $(\lambda_i, \alpha_{(i)})$ are the solutions of

$$\text{(13)} \qquad S_{xy} S_{yy}^{-1} S_{yx} \, \alpha_{(i)} = \lambda_i^2 S_{xx} \, \alpha_{(i)}$$

and

$$\beta_{(i)} = \frac{1}{\lambda_i} S_{yy}^{-1} S_{yx} \, \alpha_{(i)} \qquad .$$

To solve (13) numerical algorithms based on CD and SVD have appeared in [3] and [10] respectively. The former performs a Cholesky decomposition of S_{xx} (and S_{yy}) while checking for ill-conditioning, i.e. checking for the near-dependence of the variables X_i's (and Y_j's). The latter is more stable numerically: S_{xx} and S_{yy} are not computed, instead the X and Y data matrices, of size $n \times p$ and $n \times q$, respectively, are decomposed via Householder's method such that

$$X = Q_x \begin{pmatrix} \tilde{R}_x \\ --- \\ 0 \end{pmatrix} \text{ and } Y = Q_y \begin{pmatrix} \tilde{R}_y \\ --- \\ 0 \end{pmatrix} .$$

The canonical correlations are then the singular values of $Q_x'Q_y$.

2.2 Discriminant Analysis

The problem of discrimination [12] arises in the study of G, $G \geq 2$ a priori defined groups of populations assumed to have equal variance-covariance matrices. It involves finding linear combinations (discriminant functions) of the p variates that enable the experimenter to "best" represent the groups by maximizing among-group relative to within-group variability. The discriminant functions $d_{(i)}$, $i=1,2,\ldots,r$ where $r = \min(p,G-1)$ are usually found by solving the problem

$$\max_{d_{(i)}} \lambda^2 = \frac{d_{(i)}'Bd_{(i)}}{d_{(i)}'Td_{(i)}} = \frac{d_{(i)}'Bd_{(i)}}{d_{(i)}'(W + B)d_{(i)}} \tag{14}$$

where W, B, T denote the within groups, between groups, and total sum of squares and cross-product matrices. (14) reduces to solving

$$(B - \gamma W)\, d_{(i)} = 0 \tag{15}$$

where

$$\gamma = \frac{\lambda^2}{1 - \lambda^2}$$

is the largest eigenvalue of $W^{-1}B$. Clearly, the GSE algorithm can be used to solve (15). If we now define $q = G-1$ dummy variables Y_i, $i=1,\ldots,q$ such that

$$y_{ij} = \begin{cases} 1 & \text{if the } j^{\underline{th}} \text{ observation belongs to the } i^{\underline{th}} \text{ group} \\ 0 & \text{otherwise} \end{cases}$$

then the set of variables $X_1, X_2, \ldots, X_p$ and $Y_1, Y_2, \ldots, Y_q$ can now be used in the canonical analysis framework of §2.1. In discriminant analysis, it is usual that $q \ll p$ so that the SVD should operate of $Q_y'Q_x$ to yield the singular values λ_i , $i=1,\ldots,q$.

2.3 The Multivariate Linear Model

A multivariate linear model is of the form

(16) $$Y = XB + U$$

where Y is an $n \times q$ matrix of observed values of q variates, X is an $n \times p$ matrix of observed values of p non-stochastic variables, or is a design matrix , B is a $p \times q$ matrix of unknown parameters to be estimated and U is an $n \times q$ matrix of errors where U_i is multivariate normal $(0,\Sigma)$ and $E(U_i U_j') = 0$ for $i \neq j$. A multivariate linear hypothesis is of the form H_o: CB = 0 where C is a $p \times p$ matrix with unity on the diagonals corresponding to the columns omitted from X to yield X_o, the restricted design matrix, and zeroes elsewhere. All the tests for H_o require the formation of the matrices (see e.g. [14])

(17) $$E = Y'QY$$

and

(18) $$H = Y'(P - P_o)Y$$

where

$$Q = I - P \ , \ P = X(X'X)^{-1}X' \ ;$$

$$P_o = X_o(X_o'X_o)^{-1}X_o'$$

Several criteria for testing H_o have been put forward and they are the results of invariant transformations, see, in particular, [1]. The test statistics which are associated with them depend on the eigenvalues of

$$(19) \qquad H v = \lambda E v$$

If the test criteria are based on λ_{min} (or λ_{max}) then we can proceed as in 1.2 and compute the smallest (or largest) eigenvalue by standard methods. If the criteria involve all the eigenvalues λ's then SVD can be used with the ideas of [16] transferred to Householder's and Cholesky's decomposition.

2.4 Limited Information Maximum Likelihood

A system of stochastic equations is usually written as

$$(20) \qquad GY - FX = E$$

where Y is the $q \times n$ matrix of endogenous variables, X is the $p \times n$ matrix of exogenous variables, E is the $q \times n$ matrix of random errors, and G $q \times q$ and F $q \times p$ are both unknown matrices. See e.g. [9]

Before estimating the elements of the matrices G and F, the model must be identified. Econometric theory generally provides a priori information in that we often know that certain variables do not appear in certain equations (i.e. certain elements of G and F are restricted to be zero). Without loss of generality, consider the first equation of (20)

$$g_{(1)}Y_* - f_{(1)}X_* = e_{(1)}$$

with the elements of $g_{(1)}$ and $f_{(1)}$ suitably rearranged to correspond to the included exogenous and endogenous X_* and Y_* variables, respectively.

It is also assumed that the coefficient of the first endogenous variable is equal to 1 (normalization rule). If this particular equation is over-identified, the limited information maximum likelihood procedure amounts to

$$(21) \qquad \min_{g_{(1)}} \frac{g_{(1)}' Y_*' Q_* Y_* \, g_{(1)}}{g_{(1)}' Y_*' Q \, Y_* \, g_{(1)}}$$

where $\qquad Y_*$ is $q_* \times n$, X_* is $p_* \times n$

$$Q_* = I - X_*(X_*'X_*)^{-1}X_*'$$

$$Q = I - X(X'X)^{-1}X'$$

and (21) reduces in turn to the eigenproblem

$$(22) \qquad (Y_*'Q_*Y_* - \gamma \; Y'QY) \; g_{(1)} = 0$$

for which we choose the smallest eigenvalue γ_{min}. Then

$$(23) \qquad f_{(1)} = -(X_*'X_*)^{-1}X_*'Y_*g_{(1)}$$

CD can be used to solve (22) and extract <u>only</u> the smallest eigenvalue by standard methods. SVD cannot be used to extract <u>only</u> the smallest singular value. Rather, all singular values are computed as proposed in [4] ,where the authors work with the data matrices X and Y_* instead of cross product matrices.

The following table summarizes the four eigenvalues problems associated with the multivariate models and the test criteria which have been proposed in the literature.

SUMMARY TABLE

	General Symmetric Eigenproblem	Canonical Analysis Model	Discriminant Analysis Model	Multivariate Linear Model	Limited Information Maximum Likelihood Method
		$X_1, X_2, \ldots, X_p$ $Y_1, Y_2, \ldots, Y_q$	$X_1, X_2, \ldots, X_p$ G groups	$X_1, X_2, \ldots, X_p$ $Y_1, Y_2, \ldots, Y_q$	$X_1, X_2, \ldots, X_{p*}$ $Y_1, Y_2, \ldots, Y_{q*}$
	A	$S_{xy} S_{yy}^{-1} S_{yx}$	B: between-group sum of squares & cross-product matrix	H: "hypothesis" sum of of squares & cross-product matrix	$Y_*' Q_* Y_*$ $Q_* = I - X_*(X_*' X_*)^{-1} X_*'$
	B	S_{xx}	W: within-group sum of squares & cross-product matrix	E: "error" sum of squares & cross-product matrix	$Y'QY$ $Q = I - X(X'X)^{-1}X'$
	$Ax = \lambda Bx$	$(S_{xy} S_{yy}^{-1} S_{yx})\alpha = \rho^2 S_{xx} \alpha$	$Bd = \lambda Wd$	$Hv = \lambda Ev$	$(Y_*' Q_* Y_*) g = \gamma (Y'QY)$
	$\lambda_1, \lambda_2, \ldots, \lambda_m$	$\rho_1, \rho_2, \ldots, \rho_r$ canonical correlation $r = \min(p,q)$	$\lambda_1, \lambda_2, \ldots, \lambda_r$ discriminant criterion $r = \min(p, G-1)$	$\lambda_1, \lambda_2, \ldots, \lambda_r$ $r = \min$ (q, df for the hypothesis	$\gamma_1, \gamma_2, \ldots, \gamma_r$
Test Statistics	$\mathrm{tr}(B^{-1}A) = \sum_i \lambda_i$			$\mathrm{tr}(HE^{-1}) \equiv$ Lawley's V criterion $\Rightarrow$ Hotelling's T^2	
	$\frac{\lvert B \rvert}{\lvert A+B \rvert} = \frac{1}{\prod_i (1+\lambda_i)}$	$\prod_i^r (1-\rho_i^2) \equiv$ Wilk's Λ criterion Hotelling's coefficient of alienation	$\frac{\lvert W \rvert}{\lvert B+W \rvert} = \frac{\lvert W \rvert}{\lvert T \rvert} \equiv$ Wilk's Λ criterion $\Rightarrow$ Bartlett's V, Rao's F	$\frac{\lvert E \rvert}{\lvert H+E \rvert} \equiv$ Wilk's Λ criterion $\Rightarrow$ Bartlett's V, Rao's F	
	$\lambda_{max} \equiv \lambda_1$			Roy's criterion $\frac{\lambda_{max}}{1+\lambda_{max}}$	
	$\lambda_{min} \equiv \lambda_m$			Anderson's criterion	γ_{min}
	$\mathrm{tr}(A(A+B)^{-1}) = \Sigma\left(\frac{\lambda_i}{1+\lambda_i}\right)$			Pillai's criterion	

3. PRACTICAL CONSIDERATIONS AND CONCLUSIONS

3.1 The choice between a Cholesky decomposition for SGE and a Householder decomposition for SVD is important to note for practical implementation. Typically, the difference is between operating with symmetric p x p matrices or rectangular matrices, say n x p, $n \geq p$. In the former case, the matrices are usually sample correlation or variance-covariance matrices. In the latter case, the matrices represent the "raw" data to which Householder's transformations are applied. The number of multiplications required for a Cholesky decomposition is about $\frac{1}{6}p^3$ (plus $np^2/2$ multiplications to obtain the correlation or covariance matrix) whereas the number of multiplications for a Householder decomposition is approximately $np^2 - p^3/3$. Thus, for "tall" data matrices $n \gg p$, CD is more economical.

3.2 Storage considerations may not permit an in-core data matrix but will favor a compact storage scheme for the symmetric matrices which are needed for other statistical computations. The design philosophy of many statistical packages, e.g., SPSS [15], is to handle n x p data matrices with unlimited number of observations. The data matrix resides on disk, while the correlation matrix resides in central memory. This precludes the use of HD for all but small size problems. It is possible, of course, to perform HD on very tall matrices by annihilating rows $p + 1, p + 2, \ldots, n$, one at the time, provided that the $(p + 1) \times p$ matrix fits in core.

3.3 An important requirement of mathematical software is the reliability of its numerical algorithms. In the previous section, we have indicated how linear algebra algorithms apply to multivariate statistical models. In eigen-analysis, the researcher is fortunate to have the EISPACK [7] subroutines developed at Argonne Laboratories, and based on the pioneering work of Wilkinson and others[20] for inclusion in a statistical programming system.* There are instances when a user inputs one or more correlation or variance-covariance matrices into the statistical package. (SPSS provides this convenient feature.) These previously computed matrices can be ill-conditioned if the original variables are almost linearly dependent. Or they can fail the test for positive definiteness if missing data in the original variables lead to computing pairwise correlation coefficients with differing number of degrees of freedom in the correlation matrix. These difficulties can be detected in a Cholesky decomposition. In a recent paper [8] the problem of an ill-conditioned GSE $Ax = \lambda Bx$ has been studied for the cases when A or B or both are perturbed or are ill-conditioned, and their effect on the stability of the eigenvalues λ.

---

* At Northwestern University's Vogelback Computing Center, the algorithms for canonical correlation and discriminant analysis based on SGE have been implemented in the SPSS package for CDC 6000 machines [18].

3.4 In sum, if variance-covariance/correlation matrices are computed accurately, Cholesky decomposition/SGE is numerically stable and cheapest. If not, Householder's decomposition will avoid possible ill-conditioning and SVD will also provide a good global definition of the rank of the data matrices involved.

The purpose of this paper has been to bring to the attention of statisticians numerically acceptable procedures and encourage statistical programmers to take advantage of the advances made by numerical analysts in recent years. The existence of statistical packages has made it too easy to go to the computer for data analysis and produce sense or nonsense at an astonishing rate. If we believe that we study the right model then we certainly wish that the results are obtained with greatest working accuracy.

References

1. Anderson, T. W., Introduction to Multivariate Statistical Analysis, John Wiley & Sons, 1968.
2. Businger, P. and G. Golub, "Linear Least Squares Solutions by Householders' Transformations", Numerische Mathematik, 7, (1965), pp. 269-276.
3. Cohen, C. and A. Ben-Israel, "On the Computation of Canonical Correlations", Cahier du Centre d'Etudes de Recherche Operationelle, 11 (1969), pp. 121-132.
4. Dent, W. and G. Golub, "Computation of the Limited Information Maximum Likelihood Estimator", STAN-CS-73-339 (1973). Computer Science Department, Stanford University, Stanford, California.
5. Eckart, C. and G. Young, "The Approximation of One Matrix by Another of Lower Rank," Psychometrika, 1, (1936), pp. 211-218.
6. Eckart, C. and G. Young, "A Principal Axis Transformation for Non-Hermitian Matrices", Bull. Amer. Math. Soc., 45, (1939), pp. 118-121.
7. EISPACK Guide, Matrix Eigensystem Routines, B. T. Smith et. al., eds., Springer-Verlag (1974).
8. Fix, G. and R. Heiberger, "An Algorithm for the Ill-Conditioned Generalized Eigenvalue Problem", SIAM J. on Num. Analysis, 9, (1972), pp. 78-88.
9. Goldberger, A., Econometric Theory, John Wiley & Sons, (1964).
10. Golub, G., "Matrix Decompositions and Statistical Computations", in Statistical Computation, (Milton and Nelder, eds.), Academic Press, (1969).
11. Golub, G. and C. Reinsch, "Singular Value Decomposition and Least-Squares Solutions", Numerische Mathematik, 14, (1970), pp. 403-420.
12. Kendall, M.G., A Course in Multivariate Statistics, Griffin (1967).

13. Martin, R. S. and J. H. Wilkinson, "Reduction of the Symmetric Eigenproblem $Ax = \lambda Bx$ and Related Problems to Standard Form", Numerische Mathematik, 11, (1968), pp. 99-110.

14. Morrison, D. F., Multivariate Statistical Methods, McGraw-Hill Book Company (1967).

15. Nie, N., D. Bent, and C. D. Hull, Statistical Package for the Social Sciences, McGraw-Hill Book Company, 1970.

16. Roy, S. N. and A. E. Sarhan, "On Inverting a Class of Patterned Matrices", Biometrika, 43, (1952), pp. 227-231.

17. Schucany, W. R., B. S. Shannon, and P. D. Minton, "A Survey of Statistical Packages", ACM Computing Surveys, 4, (1972), pp. 65-80.

18. SPSS-6000, Statistical Package for the Social Sciences, Update Manual for CDC-6000 machines, Version 5.5, Vogelback Computing Center, Northwestern University, Evanston, Illinois, see also [15].

19. Wampler, R. H., "A Report on the Accuracy of Some Widely Used Least Squares Computer Programs", J. of Am. Statistical Association, 65, pp. 549-565.

20. Wilkinson, J. H. and C. Reinsch, Linear Algebra, Handbook for Automatic Computation, Volume II, Springer-Verlag (1971).