CSQ – Computational Statistics Quarterly, Vol. 1, Issue 1, 1984, page 3 – 12.

131

Fixed Rank Matrix Approximation with Singular Weights Matrices

By *J. de Leeuw*, Leiden

Summary

In this paper we approximate a matrix by another matrix of lower rank. The approximation is defined by using the general class of orthogonally invariant norms, in combination with row-weights and column-weights which can be singular. Our results generalize the existing ones.

Keywords

matrix, singular value, approximation, reduction of dimensionality

Authors adress

Jan de Leeuw

Department of Data Theory FSW/RUL

Middelstegracht 4

2312 TW Leiden

The Netherlands

1: Introduction

Suppose X is a given n x m matrix. The problem of minimizing the norm $\mathrm{tr}\,(X - Y)'(X - Y)$ over all matrices Y with $\mathrm{rank}(Y) \leqslant \rho$, and the solution to this problem are very well known. Some of the relevant references are Schmidt (1907), Eckart and Young (1936), Householder and Young (1938), Keller (1962). The application of these results to various forms of factor analysis, principal component analysis, correspondence analysis, multidimensional scaling, and other graphical data analysis techniques are much too numerous to list here. Many of these applications are reviewed by Gabriel (1971), Gnanadesikan (1977), Kruskal (1978). From these reviews it is clear that this matrix approximation result is one of the basic tools of psychometrics, and perhaps of data analysis in general.

Jan de Leeuw, Department of Data Theory, Leiden University, NL–Leiden.

The matrix approximation problem mentioned above, and its solution, have been generalized recently in various directions. The first direction is weighted least squares. This only requires a minor adaptation of the classical results. Keller and Wansbeek (1983) and Van Praag (1982) study the problem of minimizing $\mathrm{tr}\,(X - Y)'A(X - Y)$ for a positive definite matrix of row-weights A, and they discuss applications of this result to multinormal maximum likelihood theory. They both indicate that the case in which A is singular may be of interest, but they do no present definite results on this case. De Leeuw (1981) studied the case in which A is singular, using results from penalty function theory and perturbation theory. Although these are useful tools, they lead to fairly heavy computations. One of the purposes of this paper is to solve the singular case by purely algebraic methods, which turn out to be at least as powerful as the analytic ones.

Other recent generalizations steer away from the Euclidean norm. A general reference is Fiedler (1968). Maitre (1968) uses the class of generalized norms introduced into numerical analysis by Gastinel (1962). In a classical paper Mirsky (1960) uses the unitarily invariant norms introduced by Von Neumann (1937). Reviews of the application of unitarily invariant norms to matrix approximation problems are Corsten (1976) and Rao (1980). In Rao's paper there is a general result which covers the case of minimizing $||A(X - Y)B||$ for positive definite matrices A and B of row-weights and column-weights, and for an arbitrary unitarily invariant norm. This is the result we want to generalize in this paper.

For ease of reference we summarize some of the basic definitions here. Because we work with real matrices, we define orthogonally invariant norms on the space of all real n x m matrices. They must satisfy

a: $||X|| > 0$ for $X \neq \emptyset$ (the null matrix),

b: $||cX|| = |c|\cdot||X||$ for all real c,

c: $||X + Y|| \leqslant ||X|| + ||Y||$,

d: $||VXU|| = ||X||$ for all square orthonormal V and U.

Condition (d) is what makes these norms special. The singular value decomposition tells us that any matrix X can be decomposed as $X = V\Lambda U'$, with V and U square orthonormal, and with Λ pseudo-diagonal and non-negative. (A matrix is called pseudo-diagonal if all is off-diagonal elements are zero, this does not require the matrix to be square). It follows from (d) that $||X|| = ||\Lambda||$, i.e. the norm of X is a function of its singular values only. Von Neumann (1937) establishes a one-to-one correspondence between orthogonally invariant norms and symmetric gauge functions. A symmetric gauge function ϕ is a real valued function, defined on a space of real vectors, such that

a: $\phi(x) > 0$ for $x \neq \emptyset$ (the null vector),

b: $\phi(\xi x) = |\xi|\cdot\phi(x)$ for all real ξ,

c: $\phi(x + y) \leq \phi(x) + \phi(y)$,

d: $\phi(Px) = \phi(x)$ for all permutation matrices P,

e: $\phi(Sx) = \phi(x)$ for all sign matrices S.

Remember that a permutation matrix is a zero-one matrix, with exactly one element equal to one in each row and column. A sign matrix is diagonal with plus or minus one on the diagonal. Von Neumann's result is that each orthogonally invariant norm can be written in the form $||X|| = \phi(\lambda)$, where λ are the singular values Λ, collected in a vector.

It is clear that orthogonally invariant norms define a very general class, the only restriction being that they are symmetric gauges on the singular values. The Gastinel-norms, mentioned above, are a somewhat different class which can be defined in terms of certain generalized singular values. It is not true, however, that all possible cases of interest are covered by these two classes. In Bargmann and Baker (1977) for example, the function $||X - Y||$, with $||\cdot||$ the ℓ_∞-norm, is minimized over all Y with $\text{rank}(Y) \leq \rho$. The ℓ_∞-norm on the elements of the matrix is not orthogonally invariant, although the ℓ_∞-norm on the vector of singular values defines an orthogonally invariant matrix norm. Similar comments apply to fitting in the ℓ_1-norm, which is discussed briefly in Gabriel and Odoroff (1983).

It is now easy to describe the contents of this paper. We want to minimize $||A(X - Y)B'||$ over Y with $\text{rank}(Y) \leq \rho$, where A and B need not even be square, and where $||\cdot||$ is any orthogonally invariant norm. Because of the particular application in Keller and Wansbeek (1983), which provided the motivation for doing this research, we also discuss a restricted version of the problem. Thus we generalize Rao (1980), because we allow for singular row-weights and column-weights. And we generalize De Leeuw (1981), because we include orthogonally invariant norms other than the ℓ_2-norm.

2: Main approximation result

Suppose we want to approximate a given $n \times m$ matrix X with a matrix Y, also $n \times m$, and with $\text{rank}(Y) \leq \rho$. Closeness of approximation is measured by using a $p \times n$ matrix A of row-weights, a $q \times m$ matrix B of column-weights, and by defining the loss-function

$$\sigma(Y) = ||A(X - Y)B'||, \tag{1}$$

where $||\cdot||$ is an orthogonally invariant matrix norm on the space of $p \times q$ matrices.

We start with some convenient definitions. Suppose $A = P\Psi Q'$ and $B = G\Phi H'$ are singular value decompositions of A and B. Thus P, Q, G, H are square orthonormal. Both Ψ and Φ are pseudo-diagonal, Ψ is $p \times n$, Φ is $q \times m$, and we assume without loss of generality that the elements of Ψ and Φ decrease along the diagonal. Suppose $\text{rank}(A) = s$. Partition Q as $Q = (Q_1 \mid Q_0)$, where Q_1 is $n \times s$ and Q_0 is $n \times (n - s)$. In the same way $H = (H_1 \mid H_0)$, where H_1 is $m \times t$ and H_0 is $m \times (m - t)$, for $t = \text{rank}(B)$. Moreover Ψ_1 and Φ_1 are the leading $s \times s$ and $t \times t$ positive definite diagonal submatrices of Ψ and Φ (our singular values are always chosen to be non-negative). Observe that it is possible that $s = n$ and/or $t = m$. In that case Q_0 and/or H_0 have no columns, but the formulas we derive can still be interpreted in the obvious way. In fact they can even be interpreted in the (admittedly completely uninteresting) case in which $s = 0$ and/or $t = 0$.

Our first step in the construction of the optimal Y is a change of variables. Write Y in the form

$$Y = Q_1FH_1' + Q_1CH_0' + Q_0DH_1' + Q_0EH_0'. \tag{2}$$

We now want to minimize

$$\sigma(F,C,D,E) = \left|\left|\begin{matrix} \Psi_1Q_1'XH_1\Phi_1 - \Psi_1F\Phi_1 & \phi \\ \phi & \phi \end{matrix}\right|\right|, \tag{3}$$

over all F, C, D, E that satisfy

$$\text{rank}\begin{vmatrix} F & C \\ D & E \end{vmatrix} \leq \rho. \tag{4}$$

Because C, D, E do not appear on the right-hand side in (3) it is best to interpret (4) as a condition on F. Thus we require that F is such that there exist C, D, E of the appropriate size such that (4) is true. But a little reflection shows that this condition is simply equivalent to rank(F) $\leq \rho$. Thus we can find F, C, D, E by first minimizing (3) over F, under the condition that rank(F) $\leq \rho$. This gives a solution $\hat{F}$. We can then choose C, D, E arbitrarily, except for the fact that together with $\hat{F}$ they must satisfy (4).

But we know how to construct $\hat{F}$ from Mirsky (1960). If $U = \Psi_1Q_1'XH_1\Phi_1$, and $U = S\Omega T'$ is a singular value decomposition of U, then

$$\hat{F} = \Psi_1^{-1}S\{\Omega\}_\rho T'\Phi_1^{-1}. \tag{5}$$

Here $\{\Omega\}_\rho$ is pseudo-diagonal, of order s x t, with its ρ largest elements equal to those of Ω, and its other elements equal to zero. Thus $\{\Omega\}_\rho$ is the best rank-ρ approximation to Ω. Using this interpretation it also makes sense to write $\hat{F} = \Psi_1^{-1}\{U\}_\rho\Phi_1^{-1}$, with $\{U\}_\rho$ the best rank-ρ approximation to U. It is of some interest to observe that $\hat{F}$ is not necessarily uniquely defined by (5). If rank(U) $> \rho$ and $\omega_\rho = \omega_{\rho+1}$, then different choices of $\hat{F}$ are possible, because we can choose different elements from the singular subspace corresponding with the singular value ω_ρ. We collect our results so far in a theorem.

<u>Theorem 1</u>: Consider the problem of minimizing the loss function $\sigma(Y) = ||A(X - Y)B'||$ over all Y such that rank(Y) $\leq \rho$, where $||\cdot||$ is any orthogonally

invariant norm. The general solution for Y is

$$\hat{Y} = Q_1\hat{F}H_1' + Q_1\hat{C}H_0' + Q_0\hat{D}H_1' + Q_0\hat{E}H_0', \tag{6}$$

with

$$\hat{F} = \Psi_1^{-1}S\{\Omega\}_\rho T'\Phi_1^{-1}, \tag{7}$$

and with $\hat{C}$, $\hat{D}$, $\hat{E}$ chosen in such a way that

$$\text{rank}\begin{vmatrix}\hat{F} & \hat{C}\\ \hat{D} & \hat{E}\end{vmatrix} \leq \rho.$$

Proof: Given above. □

Condition (8) can be made somewhat more explicit. Suppose $\rho_0 = \text{rank}(\Omega)$ and $\rho_1 = \text{rank}\ \{\Omega\}_\rho$, i.e. $\rho_1 = \min(\rho,\rho_0)$. Moreover Ω_1 is the leading $\rho_1 \times \rho_1$ submatrix of Ω and of $\{\Omega\}_\rho$. The corresponding left and right singular vectors are in S_1 and T_1, the remaining singular vectors are in S_0 and T_0. Now

$$\text{rank}\begin{vmatrix}\hat{F} & \hat{C}\\ \hat{D} & \hat{E}\end{vmatrix} = \text{rank}\begin{vmatrix}\{\Omega\}_\rho & S'\Psi_1\hat{C}\\ \hat{D}\Phi_1 T & \hat{E}\end{vmatrix} = \text{rank}\begin{vmatrix}\Omega_1 & \emptyset & S_1'\Psi_1\hat{C}\\ \emptyset & \emptyset & S_0'\Psi_1\hat{C}\\ \hat{D}\Phi_1T_1 & \hat{D}\Phi_1T_0 & \hat{E}\end{vmatrix}. \tag{9}$$

By a familiar theorem, given for example by Guttman (1946), this gives

$$\text{rank}\begin{vmatrix}\hat{F} & \hat{C}\\ \hat{D} & \hat{E}\end{vmatrix} = \rho_1 + \text{rank}\begin{vmatrix}\emptyset & S_0'\Psi_1\hat{C}\\ \hat{D}\Phi_1T_0 & \hat{E} - \hat{D}\Phi_1T_1\Omega_1^{-1}S_1'\Psi_1\hat{C}\end{vmatrix}. \tag{10}$$

In fact we can go further. By using the methods of Meyer (1973), Marsaglia and Styan (1974), Oellette (1978), De Leeuw (1982), we can derive from (10) the result

$$\text{rank}\begin{vmatrix}\hat{F} & \hat{C}\\ \hat{D} & \hat{E}\end{vmatrix} = \rho_1 + \text{rank}(S_0'\Psi_1\hat{C}) + \text{rank}(\hat{D}\Phi_1T_0) + $$
$$+ \text{rank}\{K_0'(\hat{E} - \hat{D}\Phi_1T_1\Omega_1^{-1}S_1'\Psi_1\hat{C})L_0\}, \tag{11}$$

where K_0 is an orthogonal basis for the null space of $T_0'\Phi_1\hat{D}'$ and L_0 is an orthogonal basis for the null space of $S_0'\Psi_1\hat{C}$.

On the basis of these results we can distinguish two different cases. If $\rho_1 = \rho_0$, i.e. $\rho_0 \leq \rho$, then $\sigma(\hat{Y}) = 0$. By a suitable choice of $\hat{C}$, $\hat{D}$, $\hat{E}$ we can give $\hat{Y}$ any rank between ρ_0 and ρ. If $\rho_1 = \rho$, i.e. if $\rho \leq \rho_0$, then $\text{rank}(\hat{Y}) \leq \rho$

if and only if $S_0'\Psi_1\hat{C} = 0$ and $\hat{D}\Phi_1T_0 = 0$ and $\hat{E} = \hat{D}\Phi_1T_1\Omega_1^{-1}S_1'\Psi_1\hat{C}$. Here $\sigma(\hat{Y}) > 0$, except in the boundary case $\rho = \rho_0$.

3: Restricted approximation

We have seen in the previous section that $\sigma(\hat{Y}) = 0$ if and only if rank(U) = rank($Q_1'XH_1$) $\leq \rho$. This implies that $Q_1'\hat{Y}H_1 = Q_1'XH_1$, but if either A or B is not of full column rank it does not follow that $\hat{Y} = X$. We can have $\sigma(\hat{Y}) = 0$ and $\hat{Y} \neq X$, or, to put it differently, $||A(\cdot)B'||$ is not a norm on the space of all n x m matrices. But consider the subspace of all n x m matrices Z that satisfy $Q_1'ZH_0 = \phi$, $Q_0'ZH_1 = \phi$, and $Q_0'ZH_0 = \phi$. For any Z in this subspace we clearly have $||AZB'|| = 0$ if and only if $Z = \phi$. Or, if Y satisfies $Q_1'YH_0 = Q_1'XH_0$, $Q_0'YH_1 = Q_0'XH_1$, and $Q_0'YH_0 = Q_0'XH_0$, then we also have $\sigma(Y) = 0$ if and only if $Y = X$. In some applications, such as the multinormal maximum likelihood context of Keller and Wansbeek (1983), it is necessary to work with norms on subspaces instead of pseudo-norms on the whole space. Thus, accordingly, we now formulate the restricted approximation problem in which $\sigma(Y)$ is minimized over all Y that satisfy $Q_1'YH_0 = Q_1'XH_0$, $Q_0'YH_1 = Q_0'XH_1$, $Q_0'YH_0 = Q_0'XH_0$. By using the same reasoning as in the proof of theorem 1 we obtain a similar theorem for restricted approximation.

Theorem 2: Consider the problem of minimizing $\sigma(Y)$ over all Y such that rank(Y) $\leq \rho$ and $Q_1'YH_0 = Q_1'XH_0$, $Q_0'YH_1 = Q_0'XH_1$, $Q_0'YH_0 = Q_0'XH_0$. The general solution for Y is

$$\hat{Y} = Q_1\hat{F}H_1' + Q_1\hat{C}H_0' + Q_0\hat{D}H_1' + Q_0\hat{E}H_0', \tag{12}$$

with

$$\hat{F} = \Psi_1^{-1}S\{\Omega\}_\theta T'\Phi_1^{-1}, \tag{13}$$

$$\hat{C} = Q_1'XH_0, \tag{14}$$

$$\hat{D} = Q_0'XH_1, \tag{15}$$

$$\hat{E} = Q_0'XH_0, \tag{16}$$

and with θ the largest integer such that

$$\text{rank}\begin{vmatrix}\hat{F} & \hat{C}\\ \hat{D} & \hat{E}\end{vmatrix} \leq \rho. \tag{17}$$

Proof: As in the proof of theorem 1 the problem can be reduced to minimizing

$$\left\|\begin{matrix} \Omega - \Psi_1 S'FT\Phi_1 & \phi \\ \phi & \phi \end{matrix}\right\| \tag{18}$$

over all F that satisfy

$$\text{rank} \begin{vmatrix} F & \hat{C} \\ \hat{D} & \hat{E} \end{vmatrix} \leqslant \rho. \tag{19}$$

The difference is that now $\hat{C}$, $\hat{D}$, and $\hat{E}$ are given matrices, they can not be chosen freely any more. That $\hat{F}$ must be of the form (13) is clear from the proof of theorem 1. If it is not, then there are non-singular transformations which transform it into this form. These transformations do not change the rank, and give a smaller loss. ∎

The expression for Y in (12) can be simplified somewhat. In the first place $Q_1Q_1'XH_0H_0' + Q_0Q_0'XH_1H_1' + Q_0Q_0'XH_0H_0' = X - Q_1Q_1'XH_1H_1'$. In the second place we can write $S\Omega_\theta T'$ as $S\Omega T'\Pi_\theta$, with Π_θ a symmetric idempotent of rank θ. This gives

$$\hat{Y} = Q_1Q_1'XH_1\Phi_1\Pi_\theta\Phi_1^{-1}H_1' + X - Q_1Q_1'XH_1H_1' =$$

$$= X - Q_1Q_1'XH_1(I - \Phi_1\Pi_\theta\Phi_1^{-1})H_1'. \tag{20}$$

Of course we can also write $S\Omega_\theta T'$ as $\Xi_\theta S\Omega T'$, with Ξ_θ another symmetric idempotent of rank θ. This gives the alternative formula

$$Y = X - Q_1(I - \Psi_1\Xi_\theta\Psi_1^{-1})Q_1'XH_1H_1'. \tag{21}$$

It is also possible to make (17) somewhat more precise, along the lines of the previous section, but we have not found a satisfactory final form. Thus for the practical problem of how to find θ we have a rather unelegant solution. We start with (20) or (21), and try all values of θ between 0 and ρ_0. We keep the largest one for which rank($\hat{Y}$) $\leqslant \rho$. In this repect, however, our results can certainly be improved.

References

Bargmann, R.E. and Baker, F.D. A minimax approach to component analysis. In P.R. Krisnaiah (ed), Applications of statistics. Amsterdam, North Holland Publishing Company, 1977.

Corsten, L.C.A. Matrix approximation: a key to multivariate methods. Paper presented at the Biometric Conference, Boston, 1976.

De Leeuw, J. On a problem of Keller and Wansbeek. Unpublished paper. Department of Datatheory FSW/RUL, Leiden University, 1981.

De Leeuw, J. The null-space of a partitioned matrix. Unpublished paper. Department of Datatheory FSW/RUL, Leiden University, 1982.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. Psychometrika, 1936, 1, 211-218.

Fiedler, M. Metric problems in the space of matrices. In Programmation en Mathematiques Numériques, Colloque Internationale du CNRS, Numéro 165, Paris, Editions CNRS, 1968.

Gastinel, N. Matrices du second degré et normes générales en Analyse Numérique. Publications scientifiques et techniques du ministère de l'air. Numéro NT 110. 1962.

Gabriel, K.R. The biplot graphical display of matrices with application to principal component analysis. Biometrika, 1971, 58, 453-467.

Gabriel, K.R. and Odoroff, C.L. Resistant lower rank approximation of matrices. Paper presented at the third international symposium Data Analysis and Informatics, Versailles, 1983.

Gnanadesikan, R. Methods for the statistical data analysis of multivariate observations. New York, Wiley, 1977.

Guttman, L. Enlargement methods for computing the inverse matrix. Annals of Mathematical Statistics, 1946, 17, 336-343.

Householder, A.S. and Young, G. Matrix approximation and latent roots. American Mathematical Monthly, 1938, 45, 165-171.

Keller, J.B. Factorization of matrices by least squares. Biometrika, 1962, 49, 239-242.

Keller, W.J. and Wansbeek, T.J. Multivariate methods for quantitative and qualitative data. Journal of Econometrics, 1983, 22, 91-112.

Kruskal, J.B. Factor analysis and principal component analysis. In W.H. Kruskal and J. Tenur (eds) International Encyclopedia of Statistics, San Francisco, Freeman, 1978.

Maitre, J.F. Approximation de rang donné dans un espace de matrices. In Programmation en Mathematiques Numériques, Colloque International du CNRS, Numéro 165, Paris, Editions CNRS, 1968.

Marsaglia, G. and Styan, G.P.H. Equalities and inequalities for ranks of matrices. Linear and Multilinear Algebra, 1974, 2, 269-292.

Meyer, C.D. Generalized inverses and ranks of block matrices. SIAM Journal on Applied Mathematics, 1973, 25, 597-602.

Mirsky, L. Symmetric gauge functions and unitary invariant norms. Quarterly Journal of Mathematics, 1960, 11, 50-59.

Ouelette, D.V. Schur complements and statistics. Unpublished masters thesis. Department of Mathematics, McGill University, 1978.

Rao, C.R. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In P.R. Krishnaiah (ed) Multivariate Analysis V. Amsterdam, North Holland Publishing Company, 1980.

Schmidt, E. Zur Theorie der linearen und nicht linearen Integralgleichungen. Mathematische Annalen, 1907, 63, 433-476.

Van Praag, B.M.S. The population sample decomposition with an application to minimum distance estimators. Unpublished paper. Center for Research in Public Economics, Leiden University, 1982.

Von Neumann, J. Some matrix inequalities and metrization of matrix space. Tomsk University Review, 1937, 1, 286-299

Received March 1983
(revised October 1983)