# A Generalization of the Eckart-Young-Mirsky Matrix Approximation Theorem

G. H. Golub*
*Department of Computer Science*
*Stanford University*
*Stanford, California 94305*

Alan Hoffman
*Mathematics Science Department*
*Thomas J. Watson Research Center*

G. W. Stewart†
*Department of Computer Science*
*and Institute for Physical Science and Technology*
*University of Maryland at College Park*
*College Park, Maryland 20742*

In memory of James H. Wilkinson

Submitted by Jack Dongarra

ABSTRACT

The Eckart-Young-Mirsky theorem solves the problem of approximating a matrix by one of lower rank. However, the approximation generally differs from the original in all its elements. In this paper it is shown how to obtain a best approximation of lower rank in which a specified set of columns of the matrix remains fixed. The paper concludes with some applications of the generalization.

*This work was supported in part by the National Science Foundation under grant DCR-8412314.

†This work was supported in part by the Air Force Office of Sponsored Research under grant AFOSR-82-0078.




52 Vanderbilt Ave., New York, NY 10017 0024-3795/87/$3.50

## 1. INTRODUCTION

Let $X$ be an $n \times p$ matrix with $n \geq p$. An important problem with diverse applications in multivariate analysis is to find approximations to $X$ that are of rank not greater than a fixed integer $r$. Specifically, let $\|\cdot\|$ be a unitarily invariant matrix norm, that is, a matrix norm that satisfies

$$\|U^T X V\| = \|X\| \tag{1.1}$$

for all unitary matrices $U$ and $V$. Then we seek an approximation $\hat{X}$ to $X$ that satisfies

$$\operatorname{rank}(\hat{X}) \leq r, \tag{1.2a}$$

$$\|\hat{X} - X\| = \inf_{\operatorname{rank}(\bar{X}) \leq r} \|\bar{X} - X\|. \tag{1.2b}$$

In 1936 Eckart and Young [1] gave an elegant constructive solution to this problem for the *Frobenius norm* defined by $\|X\|_F^2 = \operatorname{trace}(X^T X)$. This construction was later shown by Mirsky [5] to solve the problem for an arbitrary unitarily invariant norm. The construction is cast in terms of what is now called the singular value decomposition of $X$. Write

$$X = U \Psi V^T, \tag{1.3}$$

where $U^T U = V^T V = I$ and

$$\Psi = \operatorname{diag}(\psi_1, \psi_2, \ldots, \psi_p) \tag{1.4}$$

with

$$\psi_1 \geq \psi_2 \geq \cdots \geq \psi_p \geq 0. \tag{1.5}$$

Set

$$\hat{\Psi} = \operatorname{diag}(\psi_1, \psi_2, \ldots, \psi_r, 0, \ldots, 0). \tag{1.6}$$

Then

$$\hat{X} = U \hat{\Psi} V^T \tag{1.7}$$

is a matrix satisfying (1.2). For the Frobenius norm, if $\mathit{rank}(X) \geq r$, then $\hat{X}$ is unique if and only if $\psi_r > \psi_{r+1}$.

The Eckart-Young construction has the drawback that $\hat{X}$ generally differs from $X$ in all its elements, which makes it unsuitable for applications in which some of the columns of $X$ are fixed. Assuming that the fixed columns are at the beginning of the matrix $X$, we are led to consider the following problem. Let

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}, \tag{1.8}$$

where $X_1$ has $k$ columns. Find a matrix $\hat{X}_2$ such that

$$\operatorname{rank}\left[\begin{pmatrix} X_1 & \hat{X}_2 \end{pmatrix}\right] \leqslant r, \tag{1.9a}$$

$$\left\|\begin{pmatrix} X_1 & \hat{X}_2 \end{pmatrix} - \begin{pmatrix} X_1 & X_2 \end{pmatrix}\right\| = \inf_{\operatorname{rank}[(X_1 \bar{X}_2)] \leqslant r} \left\|\begin{pmatrix} X_1 & \bar{X}_2 \end{pmatrix} - \begin{pmatrix} X_1 & X_2 \end{pmatrix}\right\|. \tag{1.9b}$$

In other words, find a best rank $r$ approximation to $X$ that leaves $X_1$ fixed. This problem will be solved in the next section. In Section 3 we shall make a number of observations about the solution.

## 2. THE GENERALIZATION

Our main result is contained in the following theorem. In it we denote by $\mathbf{H}_r$ the operator that maps $X$ onto $\hat{X}$ defined by (1.7), with the convention that if $r$ is greater than the number of columns of $X$ then $\mathbf{H}_r$ is the identity.

THEOREM. *Let $X$ be partitioned as in* (1.8) *where $X_1$ has $k$ columns, and let $l = \operatorname{rank}(X_1)$. Let $P$ denote the orthogonal projection onto the column space of $X$ and $P^{\perp}$ the orthogonal projection onto its orthogonal complement. If*

$$l \leqslant r, \tag{2.1}$$

*then the matrix*

$$\hat{X}_2 = PX_2 + \mathbf{H}_{r-l}\left(P^{\perp}X_2\right) \tag{2.2}$$

*satisfies* (1.9).

*Proof.* Without loss of generality we may assume that $k = l$; for otherwise we can replace $X_1$ with a matrix having $l$ independent columns selected from $X_1$, apply the theorem, and then restore the missing columns without changing either the rank of the result or the norm of the difference.

The proof is based on the $QR$ decomposition of $X$. Specifically, there is an orthogonal matrix $Q = (Q_1\ Q_2\ Q_3)$, with $Q_1$, $Q_2$, and $Q_3$ having respectively $k$, $p-k$, and $n-p$ columns, such that

$$Q^T X = \begin{bmatrix} Q_1^T \\ Q_2^T \\ Q_3^T \end{bmatrix} (X_1 \quad X_2) = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & 0 \end{bmatrix}, \tag{2.3}$$

where $R_{11}$ and $R_{22}$ are upper triangular (for more details see [4]). Since we are dealing with unitarily invariant norms and since premultiplying a matrix by another matrix does not mix up its columns, we may solve the approximation problem for the right hand side of (2.3) and transform back to the original problem.

An elementary compactness argument shows that approximations with the required properties exist. Let

$$\begin{bmatrix} R_{11} & \hat{R}_{12} \\ 0 & \hat{R}_{22} \\ 0 & \hat{R}_{32} \end{bmatrix} \tag{2.4}$$

be one such approximation.

First observe that $R_{11}$ is nonsingular, since $X_1$ is of rank $k$. This implies that the choice of $\hat{R}_{12}$ cannot affect the rank, since we can use $R_{11}$ to eliminate it without changing $\hat{R}_{22}$ and $\hat{R}_{32}$. The following argument shows that we may take $\hat{R}_{12} = R_{12}$.

Let

$$\begin{bmatrix} E_{12} \\ E_{22} \\ E_{32} \end{bmatrix} = \begin{bmatrix} \hat{R}_{12} - R_{12} \\ \hat{R}_{22} - R_{22} \\ \hat{R}_{32} - R_{32} \end{bmatrix} \tag{2.5}$$

be the error matrix associated with the approximation. The squares of the singular values of $(E_{12}^T\ E_{22}^T\ E_{32}^T)$ are the eigenvalues of $E_{12}^T E_{12} + E_{22}^T E_{22} + E_{32}^T E_{32}$. Since $E_{12}^T E_{12}$ is positive semidefinite, these eigenvalues are not less

than the corresponding eigenvalues of $E_{22}^T E_{22} + E_{32}^T E_{32}$. It then follows from results in [5] that

$$\left\| \begin{pmatrix} E_{12}^T & E_{22}^T & E_{32}^T \end{pmatrix} \right\| \geq \left\| \begin{pmatrix} 0 & E_{22}^T & E_{32}^T \end{pmatrix} \right\|; \tag{2.6}$$

that is, independently of the choices of $\hat{R}_{22}$ and $\hat{R}_{32}$, the choice $\hat{R}_{12} = R_{12}$ minimizes the error.

With this choice of $\hat{R}_{12}$, we now seek the approximations $\hat{R}_{22}$ and $\hat{R}_{32}$. Since $R_{11}$ is nonsingular, $Q^T X$ will have rank less than or equal to $r$ if and only if $(\hat{R}_{22}^T \ \hat{R}_{32}^T)^T$ has rank $r - l$. But a best rank $r - l$ approximation to $(R_{22}^T \ R_{32}^T)^T$ is given by

$$\begin{bmatrix} \hat{R}_{22} \\ \hat{R}_{32} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{r-l}(R_{22}) \\ 0 \end{bmatrix}. \tag{2.7}$$

Transforming back to the original problem, we get the approximation

$$\hat{X}_2 = Q_1 R_{12} + Q_2 \mathbf{H}_{r-l}(R_{22}) = Q_1 R_{12} + \mathbf{H}_{r-l}(Q_2 \mathbf{R}_{22}). \tag{2.8}$$

It now remains only to observe that $Q_1 R_{12} = P X_2$ and $Q_2 R_{22} = P^{\perp} X_2$. ■

## 3. COMMENTS

In this section we shall survey some consequences of the result established in Section 2. Except for the uniqueness condition below, we shall assume that $X_1$ has full column rank.

*Uniqueness*

The matrix $\hat{X}_2$ is unique if and only if $\mathbf{H}_{r-l}(P^{\perp} X_2)$ is unique. For the Frobenius norm, this means that $\hat{X}_2$ is unique if and only if the $(r - l)$th singular value of $P^{\perp} X_2$ is strictly greater than the $(r - l + 1)$th.

*Useful Formulas*

The relation

$$\mathbf{H}_{r-l}(Q_2 R_{22}) = Q_2 \mathbf{H}_{r-l}(R_{22}), \tag{3.1}$$

which was used in (2.8), has the computational consequence that to de-

termine $\hat{X}_2$ we need only compute the singular value decomposition of $R_{22}$, a matrix which is smaller that $P^{\perp}X_2$. Since $R_{22}$ is the only part of the $QR$ factorization of $X$ that is altered in the passage of $\hat{X}$, it follows that

$$\|\hat{X} - X\|_F^2 = \|\mathbf{H}_{r-l}(R_{22}) - R_{22}\|_F^2, \tag{3.2}$$

which is the sum of squares of the last $p - r$ singular values of $R_{22}$.

Equation (3.2) can be cast in a more familiar form by introducing the cross-product matrix $A = X^TX$ and partitioning it conformally with (2.3):

$$A = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix}. \tag{3.3}$$

It can be shown that $R_{22}^TR_{22}$ is the *Schur complement* [6] of $A_{11}$ in $A$:

$$R_{22}^TR_{22} = A_{22} - A_{21}A_{11}^{-1}A_{21}^T. \tag{3.4}$$

Since the eigenvalues of $R_{22}^TR_{22}$ are the squares of the singular values of $R_{22}$, Equation (3.2) may be summarized by saying *the square of the distance in the Frobenius norm from $\hat{X}$ to $X$ is the sum of the last $p - r$ eigenvalues of the Schur complement of $A_{11}$ in $A$.*

*A Variational Characterization*

The sum of squares of the last $p - r$ singular values of $X$ may be written in the variational form

$$\min_{\substack{U \in \mathbf{U}^{n\times(p-r)} \\ \mathbf{R}(U) \subset \mathbf{R}(X)}} \|U^TX\|_F^2, \tag{3.5}$$

where $\mathbf{U}^{n\times(p-r)}$ denotes the space of all $n \times (p - r)$ matrices with orthonormal columns and $\mathbf{R}(X)$ is the column space of $X$. The square of the minimizing norm in (1.9) can similarly be written

$$\min_{\substack{U \in \mathbf{U}^{n\times(p-r)} \\ \mathbf{R}(U) \subset \mathbf{R}(X) \\ X_1^TU = 0}} \|U^TX\|_F^2, \tag{3.6}$$

since if $U$ is constrained to satisfy $X_1^TU = 0$ then

$$U^TX = \begin{pmatrix} U^TX_1 & U^TX_2 \end{pmatrix} = \begin{pmatrix} 0 & U^TP^{\perp}X_2 \end{pmatrix}, \tag{3.7}$$

and by (3.5) the minimum of $\|U^T P^{\perp} X_2\|_F^2$ is the sum of the squares of the last $p - r$ singular values of $P^{\perp} X_2$.

It is instructive to examine what happens when the condition $X_1^T U = 0$ is replaced by $C^T U = 0$, where $C$ satisfies $\mathbf{R}(C) \subset \mathbf{R}(X)$. In this case, we can find an orthogonal matrix $V = (V_1 \; V_2)$ such that if we set

$$Y = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix} = X \begin{pmatrix} V_1 & V_2 \end{pmatrix} \tag{3.8}$$

then $\mathbf{R}(Y_1) = \mathbf{R}(C)$. Since the condition $C^T U = 0$ is equivalent $Y_1^T U = 0$, we see that

$$\min_{\substack{U \in \mathbf{U}^{n \times (p-r)} \\ \mathbf{R}(U) \subset \mathbf{R}(X) \\ C^T U = 0}} \|U^T X\|_F^2 \tag{3.9}$$

is the square of the norm of the difference to the minimizing rank $r$ approximation to $Y$ with the columns of $Y_1$ held constant. In other words, $C$ selects a subspace on which $X$ is to remain constant.

*Centering and Collinearity Diagnostics*

The Eckart-Young theorem is the basis for examining the smallest singular value $\psi_p$ of a regression matrix $X$ to diagnose collinearity: if $\psi_p$ is small, then $X$ is very near a collinear matrix. However, this procedure is inappropriate for problems with a constant term, in which the regression matrix has the form $X = (\underline{1} \; X_2)$, where $\underline{1}$ is the vector of ones. According to our theorem, the proper approach is to project $X_2$ onto the space orthogonal to $\underline{1}$ and examine the smallest singular value of the result. Since this projection is $X_2$ with its column means subtracted out, the theorem provides another rationale for the common practice of centering regression problems with a constant term.

*Multiple Correlations and Variance Inflation Factors*

When in (1.8) the number of columns $k$ of $X_1$ is equal to the rank $r$ of the target matrix $\hat{X}$, then the construction in Section 2 simply sets $R_{22}$ to zero. In this case the square of the distance between $\hat{X}$ and $X$ becomes

$$\|R_{22}\|_F^2 = \operatorname{trace}\left(A_{22} - A_{21} A_{11}^{-1} A_{21}^T\right). \tag{3.10}$$

When $X$ has been centered and scaled so that its column norms are one, the diagonal elements of (3.4) are the multiple correlation coefficients of the

columns of $X_2$ with respect to the columns of $X_1$. Thus (3.10) provides a new interpretation of these numbers: *the sum of the multiple correlation coefficients of $X_2$ with respect to $X_1$ is the square of the norm of the smallest perturbation in $X_2$ that will make it a linear combination of the columns of $X_1$.*

When $k = r = p - 1$, so that the concern is with a perturbation in the last column alone, the matrix $R_{22}$ reduces to the $(p, p)$ element $r_{pp}$ of $R$. From the fact that $A = R^T R$ and the triangularity of $R$ it is easy to verify that when $X$ is not collinear,

$$r_{pp}^{-2} = a_{pp}^{(-1)}, \tag{3.11}$$

where $a_{pp}^{(-1)}$ denotes the $(p, p)$ element of the inverse cross-product matrix $A^{-1}$. This number has been called a variance inflation factor because, in the usual regression model, it measures the amount by which the variance in the response vector is magnified in the $p$th regression coefficient. Equations (3.4) and (3.11) show that its reciprocal is the square of the smallest perturbation in the $p$th column of $X$ that will make $X$ collinear. Since variance inflation factors are not changed by reordering the columns of $X$, we see that *the reciprocal of the $j$th variance inflation factor is the square of the norm of the smallest perturbation in the $j$th column of $X$ that will make $X$ collinear.*

*Total Least Squares*

The Gaussian regression model starts with the exact relation

$$y = Xb, \tag{3.12}$$

from which $b$ can be computed in the form

$$b = X^{\dagger} y, \tag{3.13}$$

where $X^{\dagger} = (X^T X)^{-1} X^T$ is the pseudoinverse of $X$. It is further assumed that $y$ cannot be observed; instead we observe

$$\tilde{y} = y + e. \tag{3.14}$$

Gauss [2] showed that if the elements of $e$ are uncorrelated random variables with mean zero and common variance $\sigma^2$, then the natural generalization of (3.13), i.e.

$$\hat{b} = X^{\dagger} \tilde{y}, \tag{3.15}$$

gives an optimal estimate of $\hat{b}$ in the sense that among all linear estimators satisfying

$$e = 0 \quad \Rightarrow \quad b = \hat{b}, \tag{3.16}$$

$X^{\dagger}\tilde{y}$ has the smallest variance.

Now let us suppose that $X$ is also unobserved; instead we are given

$$\tilde{X} = X + E. \tag{3.17}$$

The problem is to find a plausible way to estimate $b$.

Golub and Van Loan [3] have observed that

$$\hat{b} = X^{\dagger}(\tilde{y} - \hat{e}), \tag{3.18}$$

where $\hat{e}$ is the unique vector satisfying

$$\begin{aligned} &\text{minimize} \quad \|\hat{e}\|_F \\ &\text{subject to} \quad \operatorname{rank}\left[\begin{pmatrix} X & \tilde{y} - \hat{e} \end{pmatrix}\right] = p. \end{aligned} \tag{3.19}$$

As a generalization of this, they propose to estimate $b$ by

$$\hat{b}_{\mathrm{TLS}} = (\tilde{X} - \hat{E})^{\dagger}(\tilde{y} - \hat{e}), \tag{3.20}$$

where $\hat{E}$ and $\hat{e}$ satisfy

$$\begin{aligned} &\text{minimize} \quad \left\|\begin{pmatrix} \hat{E} & \hat{e} \end{pmatrix}\right\|_F \\ &\text{subject to} \quad \operatorname{rank}\left[\begin{pmatrix} \tilde{X} - \hat{E} & \tilde{y} - \hat{e} \end{pmatrix}\right] = p. \end{aligned} \tag{3.21}$$

In other words $\hat{E}$ and $\hat{e}$ are just the residuals from the Eckart-Young projection of $(X \;\; y)$ onto the space of matrices of rank $p$. This method, which has also been introduced from another point of view by Webster, Gunst, and Mason [9], is known as latent root regression or principle component regression to statisticians and as total least squares to numerical analysts.

The procedure does not make much sense unless the elements of $E$ and components of $e$ are independently derived and equilibrated. In statistical terms, they must be uncorrelated with mean zero and all have the same variance [7]. Sometimes this can be accomplished by scaling rows and

columns of $X$ and $y$. For example, suppose that the elements of $E$ and $e$ are uncorrelated, but the components of the $j$th column of $E$ have variance $\sigma_j$ while those of $e$ have variance $\sigma^2$. If we set

$$\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p, \sigma), \tag{3.22}$$

then $(\tilde{X}\ \tilde{y})\Sigma^{-1}$ has an error structure suitable for total least squares estimation.

This scaling procedure breaks down when some of the columns of $E$ are zero, since in this case $\Sigma$ is singular. In this case it is natural to require that the corresponding columns of $X$ be unperturbed by $\hat{E}$, since they are known exactly. If we set $\Sigma^\dagger = \operatorname{diag}(\sigma_1^\dagger, \ldots, \sigma_p^\dagger, \sigma^\dagger)$, where $\sigma^\dagger = \sigma^{-1}$ if $\nu \neq 0$ and is otherwise zero, then we may characterize $\hat{E}$ as satisfying

$$\begin{aligned} &\text{minimize} \quad \|(\hat{E} \quad \hat{e})\Sigma^\dagger\|_F \\ &\text{subject to} \quad \operatorname{rank}\left[\left(\tilde{X} - \hat{E} \quad \tilde{y} - \hat{e}\right)\right] = p. \end{aligned} \tag{3.23}$$

This is seen to be the generalized Eckart-Young projection of $(\tilde{X}\ \tilde{y})$, with its errors equilibrated onto the set of matrices of rank $p$ subject to the constraint that the columns of $X$ that are without error remain unperturbed. Thus our theorem provides a generalization of the total least squares estimate.

It is worth nothing that when all the columns of $E$ are zero, the estimate reduces to the ordinary least squares estimate. Thus our theorem embraces total least squares at one extreme (none of the columns of $E$ zero) and least squares at the other (all of the column of $E$ zero).

The estimate (3.23) can be derived independently by considering the limiting case as some of the $\sigma_j$ approach zero. In fact, asymptotic expansions in [8] can be used to show that as some of the $\sigma_j$ approach zero the total least squares estimate (3.21) approaches the estimate (3.23).

## REFERENCES

1 G. Eckart and G. Young, The approximation of one matrix by another of lower rank, *Psychometrika* 1: 211–218 (1936).

2 C. F. Gauss, Theroria combinationis observationum erroribus minimus obnoxiae, in *Werke IV*, Koniglichen Gessellschaft der Wissenschaften zu Göttingen, 1821, pp. 1–26.

3 G. H. Golub and C. Van Loan, An analysis of the total least squares problem, *SIAM. Numer. Anal.* 17:883–893 (1980).

4 ____, *Matrix Computations*, Johns Hopkins, Baltimore, 1983.

5 L. Mirsky, Symmetric gauge functions and unitarily invariant norms, *Quart. J. Math. Oxford* 11:50–59 (1960).
6 D.V. Ouellette, Schur complement and statistics, *Linear Algebra Appl.* 36:187–295 (1981).
7 G. W. Stewart, A nonlinear version of Gauss's minimum variance theorem with applications to an errors-in-the-variables model, Computer Science Technical Report TR-1263, Univ. of Maryland, 1983.
8 ____, On the asymptotic behavior of scaled singular value and $QR$ decompositions, *Math. Comp.* 43:483–489.
9 J. Webster, R. Gunst, and R. Mason, Latent root regression analysis, *Technometrics* 16:513–522 (1974).

*Received 21 January 1986; revised 7 May 1986*