CHAPTER III

MULTICOLLINEARITY AND REDUCED-RANK ESTIMATION

3.1 Introduction

Multicollinearity has been a topic of concern in econometrics ever since the publication of Frisch's monograph (1934). Two approaches have already been discussed in the previous chapter (sections 2.7 and 2.8). In this chapter attention will revolve around the singular-value decomposition of a matrix; in the case of an $n \times k$ observation matrix X, its singular values are the positive square roots of the eigenvalues of $X'X$. If some of these are very small (and because of rounding error, a computer cannot easily distinguish between "small" and zero), classical methods of computing least-squares estimates (e.g., the Gauss-Seidel procedure) tend to be highly inaccurate. The method of computing the singular-value decomposition and replacing small singular values by zeros produces much more reliable results. Interestingly enough, statistical theory reaches a similar conclusion: replacing small singular values by zeros (which amounts to approximating X by an $n \times k$ matrix $X_{(l)}$ of reduced rank, l) leads to estimators with lower mean-square error. This theory is the subject of the present chapter.

3.2 Singular-value decomposition of a matrix

The concept of a singular-value decomposition of a matrix was introduced (independently) by Eckart & Young (1939, pp. 118–121) and Mirsky (1960).

DEFINITION 3.2.2. *Let X be any $n \times k$ matrix. If a triple (s, p, q) exists, where s is a nonnegative scalar, p is a $k \times 1$ vector, and q is an $n \times 1$ vector, such that*

$$Xp = sq \quad \text{and} \quad q'X = sp',$$

then s is called a singular value of X, and p and q are respectively called right and left singular vectors of X.

DEFINITION 3.2.2. *An $n \times k$ matrix $D = [d_{ij}]$ is called a (rectangular) diagonal matrix if $d_{ij} = 0$ for $i \neq j$.*

THEOREM 3.2.1. *Let X be any $n \times k$ matrix of rank r, and define $m = \min(n, k)$. Then there exists an $n \times n$ orthogonal matrix Q, a $k \times k$ orthogonal matrix P, and an $n \times k$ diagonal matrix D, with diagonal elements $s_1 \geqq s_2 \geqq \ldots \geqq s_m \geqq 0$ (the singular values of X), such that*

$$X = QDP', \tag{3.2.1}$$

where the columns of Q and P are respectively the corresponding unit-length left and right singular vectors of X. These singular values are unique.

PROOF. Since the $k \times k$ matrix $X'X$ is symmetric, there exists an orthogonal $k \times k$ matrix P such that

$$P'X'XP = \Lambda = \operatorname{diag}\{\lambda_i\} \tag{3.2.2}$$

where the λ_i's are the k (nonnegative) eigenvalues of $X'X$, and the $k \times 1$ columns p^i of P (the eigenvectors of $X'X$) may be ordered so that $\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_m \geqq 0$. We define $s_i = +\sqrt{\lambda_i}$ for $i = 1, 2, \ldots, m$.

Since $r = \operatorname{rank} X$, $s_r > 0$ and $s_{r+1} = \ldots = s_m = 0$. Define

$$q^i = Xp^i/s_i \quad \text{for } i = 1, 2, \ldots, r. \tag{3.2.3}$$

Then for $i, j = 1, 2, \ldots, r$, we have from (3.2.3) and (3.2.2)

$$q^i \cdot q^j = \frac{p^{i\prime}X'Xp^j}{s_i s_j} = \delta_{ij}$$

(where δ_{ij} is the Kronecker delta). For $i = r+1, \ldots, n$, choose q^i so that the columns of

$$Q = [q^1, q^2, \ldots, q^r, q^{r+1}, \ldots, q^n] \tag{3.2.4}$$

form an orthonormal set; then $Q'Q = QQ' = I_n$. From (3.2.2) we have clearly

$$Xp^i = 0 = s_i q^i \quad \text{for } i = r+1, \ldots, m. \tag{3.2.5}$$

Thus, from (3.2.3) and (3.2.5) we have

$$\begin{aligned} X = XPP' &= X[p^1, p^2, \ldots, p^k] \begin{bmatrix} p^{1\prime} \\ p^{2\prime} \\ \vdots \\ p^{k\prime} \end{bmatrix} \\ &= \sum_{i=1}^{k} Xp^i p^{i\prime} = \sum_{i=1}^{m} q^i s_i p^{i\prime} \quad (\text{since } Xp^i = 0 \text{ for } m < i < k) \\ &= [q^1, q^2, \ldots, q^m] \begin{bmatrix} s_1 & 0 & \ldots & 0 \\ 0 & s_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & s_m \end{bmatrix} \begin{bmatrix} p^{1\prime} \\ p^{2\prime} \\ \vdots \\ p^{m\prime}. \end{bmatrix} \end{aligned} \tag{3.2.6}$$

Denoting the square diagonal matrix in (3.2.6) as $S = \operatorname{diag}\{s_i\}$, (3.2.6) is equivalent to (3.2.1), where $D' = [S, 0]$ for $n \geqq k = m$ and $D = [S, 0]$ for $m = n \geqq k$.

Since, by (3.2.2), the singular values s_i of X coincide with the nonnegative square roots of the eigenvalues λ_i of $X'X$, for $i = 1, 2, \ldots, m$ (the remaining eigenvalues in the case $n < k$ being all zero), and these eigenvalues are unique, therefore, the singular values are unique. □

Note that if $n > k$ (hence $m = k$) and we partition Q as

$$Q = [Q_1; Q_2] = [q^1, q^2, \ldots, q^m; q^{m+1}, \ldots, q^n], \tag{3.2.7}$$

we may write (3.2.6) or (3.2.1) in the form

$$X = Q_1 S P'. \tag{3.2.8}$$

Likewise, if $n < k$ (hence $m = n$) and we partition P as

$$P = [P_1; P_2] = [p^1, p^2, \ldots, p^m; p^{m+1}, \ldots, p^k], \tag{3.2.9}$$

we may write (3.2.6) or (3.2.1) in the form

$$X = Q S P_1'. \tag{3.2.10}$$

Since one normally has $n > k$, the form (3.2.8) will frequently be used in applications.

We may use Theorem 3.2.1 as a way to define the Moore-Penrose generalized inverse of a matrix (Definition 2.3.1). First, we define the Moore-Penrose generalized inverse of a rectangular diagonal matrix, as follows. If D is an $n \times k$ diagonal matrix with diagonal elements $d_{ii} = s_i$, then defining

$$s_i^\dagger = \begin{cases} 1/s_i & \text{if } s_i \neq 0, \\ 0 & \text{if } s_i = 0, \end{cases} \tag{3.2.11}$$

the generalized inverse $D^\dagger$ is the $k \times n$ diagonal matrix in which each diagonal element in the transpose of D is replaced by its generalized inverse as defined by (3.2.11). Thus, denoting $S^\dagger = \operatorname{diag}\{s_i^\dagger\}$, if $n > k$ we have $D' = [S', 0]$ and $D^\dagger = [S^\dagger, 0]$. The Moore-Penrose generalized inverse of X is then given by

$$X^\dagger = P D^\dagger Q'. \tag{3.2.12}$$

It is easily verified that $X^\dagger$ and $D^\dagger$ satisfy the four properties of Definition 2.3.1. From (3.2.12) it is clear that once the singular-value decomposition of a matrix X has been computed, computation of its Moore-Penrose generalized inverse $X^\dagger$ is trivial. Computation of the singular-value decomposition, however, is far from trivial (see Golub & Kahan 1965, Golub & Reinsch 1970, Noble 1976, and Golub & Van Loan 1983).

Since one will often want to compute an oblique generalized inverse of X, satisfying property (iii′) of Definition 2.3.3, one may proceed as follows. Define

$$\dot{X} = V^{-1/2} X, \quad \text{hence } X = V^{1/2} \dot{X}, \tag{3.2.13}$$

and

$$X^\ddagger = \dot{X}^\dagger V^{-1/2}. \tag{3.2.14}$$

We verify that

$$X X^\ddagger X = V^{1/2} X | \dot{X}^\dagger V^{-1/2} | V^{1/2} \dot{X} = V^{1/2} \dot{X} \dot{X}^\dagger \dot{X} = V^{1/2} \dot{X} = X$$

(where the symbols | are inserted to facilitate checking the substitutions) and

$$X X^\ddagger V = V^{1/2} \dot{X} | \dot{X}^\dagger V^{-1/2} | V = V^{1/2} \dot{X} \dot{X}^\dagger V^{1/2}$$

which is symmetric since $V^{1/2}$ and $\dot{X}\dot{X}^\dagger$ are symmetric.

3.3 The condition number of a matrix

The condition number of a matrix is a numerical measure of the degree to which it is ill-conditioned, i.e., "close" to having deficient rank. To develop this concept we first introduce some definitions (cf. Golub & Van Loan 1983, pp. 12–14):

DEFINITION 3.3.1. *The Hölder p-norm of an $n \times 1$ vector x is defined by*

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i\|^p \right)^{1/p} \quad \text{for } p \geqq 1. \tag{3.3.1}$$

In the special case $p = 2$ this reduces to the Euclidean norm

$$\|x\|_2 \equiv \|x\| = \sqrt{x'x}. \tag{3.3.2}$$

Definition 3.3.1 is generalized to a matrix by

DEFINITION 3.3.2. *The Hölder p-norm of an $n \times k$ matrix X is defined by*

$$\|X\|_p = \sup_{b \neq 0} \frac{\|Xb\|_p}{\|b\|_p}, \tag{3.3.3}$$

where b is of order $k \times 1$. In the special case $p = 2$ this may be written

$$\|X\|_2 = \sup_{b \neq 0} \left(\frac{b'X'Xb}{b'b} \right)^{1/2}. \tag{3.3.4}$$

Note that if $k = 1$ and $X = x$, then b is scalar and (3.3.3) reduces to (3.3.1):

$$\sup_{b \neq 0} \frac{\|xb\|_p}{\|b\|_p} = \sup_{b \neq 0} \frac{|b| \, \|x\|_p}{|b|} = \|x\|_p.$$

Likewise, (3.3.4) reduces to (3.3.2).

The following is a special case of the Courant-Fischer min-max theorem (cf. Courant 1922, Fischer 1905; Bellman 1960, pp. 110–115):

LEMMA 3.3.1. *Let M be a $k \times k$ symmetric nonnegative-definite matrix with eigenvalues given by $\Lambda = diag\{\lambda_i\}$ arranged in descending order $\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_k$, where $P'MP = \Lambda$, $P'P = I$. Then*

$$\lambda_1 = \max_b \frac{b'Mb}{b'b} \quad \textit{and} \quad \lambda_k = \min_b \frac{b'Mb}{b'b}. \tag{3.3.5}$$

PROOF. Defining $c = P'b$ we have

$$\frac{b'Mb}{b'b} = \frac{c'P'MPc}{c'P'Pc} = \frac{c'\Lambda c}{c'c} = \frac{\sum_{i=1}^{k} \lambda_i c_i^2}{\sum_{i=1}^{k} c_i^2}.$$

Since

$$\sum_{i=1}^{k} \lambda_i c_i^2 \leqq \lambda_1 \sum_{i=1}^{k} c_i^2 \quad \text{and} \quad \sum_{i=1}^{k} \lambda_i c_i^2 \geqq \lambda_k \sum_{i=1}^{k} c_i^2$$

it follows that

$$\lambda_k \leqq \frac{\sum_{i=1}^{k} \lambda_i c_i^2}{\sum_{i=1}^{k} c_i^2} \leqq \lambda_1. \tag{3.3.6}$$

Since the equalities in (3.3.6) are attained by choice of $c = (0, \ldots, 0, 1)'$ and $c = (1, 0, \ldots, 0)'$ respectively, the result follows. □

LEMMA 3.3.2. *Let X be an $n \times k$ matrix of rank m, whose positive singular values are $s_1 \geqq s_2 \geqq \ldots \geqq s_m > 0$. Then*

$$\|X\|_2 = s_1.$$

PROOF. Defining $M = X'X$ and $\lambda_i = s_i^2$, where $X = QDP'$, this follows directly from Lemma 3.3.1. □

LEMMA 3.3.3. *Let X be an $n \times k$ matrix of rank m, whose positive singular values are $s_1 \geqq s_2 \geqq \ldots \geqq s_m > 0$. Then*

$$\|X^\dagger\|_2 = 1/s_m.$$

PROOF. From the singular-value decomposition $X = QDP'$ of X we have that of $X^\dagger = PD^\dagger Q'$, hence the singular values of $X^\dagger$ are $1/s_1 \leqq 1/s_2 \leqq \ldots \leqq 1/s_m$. The result then follows from Lemma 3.3.1. □

DEFINITION 3.3.3. *The condition number of an $n \times k$ matrix X of rank m is, in terms of the Hölder matrix 2-norm,*

$$\kappa(X) = \|X\|_2 \, \|X^\dagger\|_2.$$

From this definition and Lemmas 3.3.2 and 3.3.3 it follows immediately that $\kappa(X) = s_1/s_m$, i.e., the condition number of X is the ratio between its largest and its smallest positive singular value. Cf., e.g., Noble (1976, pp. 279, 295), Golub & Van Loan (1983, p. 140).

3.4 The Eckart-Young theorem

Eckart and Young (1936) furnished an algorithm for obtaining a best approximation of an $n \times k$ matrix X by an $n \times k$ matrix of rank less than that of X. In order to furnish a precise meaning to "best approximation" it is necessary to define a concept of distance between two $n \times k$ matrices.

DEFINITION 3.4.1. *The Frobenius norm of an $n \times k$ matrix X is defined as*

$$\|X\|_F \equiv \|X\| = \sqrt{tr(X'X)}.$$

It is equal to the square root of the sum of squares of all the elements of X. If X^1 and X^2 are two $n \times k$ matrices, we define the Frobenius distance between X^1 and X^2 as the Frobenius norm of their difference, i.e., $\|X^1 - X^2\|$.

DEFINITION 3.4.2. *Let X be any $n \times k$ matrix, and let Q and P respectively be $n \times n$ and $k \times k$ orthogonal matrices. Then a norm $\|X\|$ of X is said to be orthogonally invariant if it has the property $\|Q'XP\| = \|X\|$.*

The following is a simple extension of Lemma 1.1.1.

LEMMA 3.4.1. *The Frobenius norm is orthogonally invariant.*

PROOF. From Lemma 1.1.1,

$$\begin{aligned}\|Q'XP\| &= \sqrt{\mathrm{tr}(P'X'QQ'XP)} \\ &= \sqrt{\mathrm{tr}(P'X'XP)} \\ &= \sqrt{\mathrm{tr}(X'XPP')} \\ &= \sqrt{\mathrm{tr}(X'X)} \\ &= \|X\|. \quad \square\end{aligned}$$

It is readily seen that the Hölder matrix 2-norm is also orthogonally invariant.

DEFINITION 3.4.3. *We denote by $\mathcal{X}$ the set of all real $n \times k$ matrices, and by $\mathcal{X}_l$ the subset of $\mathcal{X}$ consisting of $n \times k$ matrices of rank $\leqq l$.*

In trying to approximate an $n \times k$ matrix of rank $> l$ by one of rank l, a basic problem is that the set of $n \times k$ matrices of rank exactly l is not closed. Such a set is defined by the condition that all minors (subdeterminants) of X of order greater than l vanish, and at least one minor of order l be nonvanishing. One wishes to find in this set a matrix that is closest to X in the Frobenius norm; but since this set is obviously not closed, the existence of such a matrix is not at all obvious. The procedure that is followed, therefore, is to deal with the set $\mathcal{X}_l$ of all $n \times k$ matrices X of rank $\leqq l$; this is a closed set, and can be compactified, hence a matrix in this set exists that is closest to X (see Lemma 3.4.2 below); but on the face of it, it might have rank $< l$ (see Figure 3.4.1). It has to be shown that it has rank exactly l; this is done in Theorem 3.4.1. The Eckart-Young theorem then provides the algorithm by which this matrix is determined, namely the replacement of all but the l largest singular values of X by zeros.

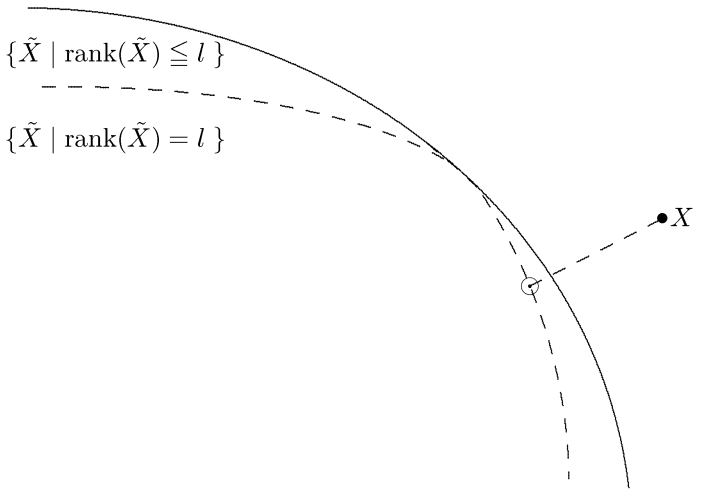

FIGURE 3.4.1

As a preliminary to the Eckart-Young theorem (Theorem 3.4.2 below), we will need Lemmas 3.4.2 and 3.4.3 and Theorem 3.4.1.

LEMMA 3.4.2. *Let X be a given $n \times k$ matrix of rank $> l$, where $l < m = \min(n, k)$. Then within the class $\mathcal{X}_l$ of $n \times k$ matrices $\tilde{X}$ of rank $\leqq l$, there exists a matrix $\hat{X}$ closest to X in the Frobenius norm, i.e., such that*

$$\|X - \hat{X}\| = \min_{\tilde{X} \in \mathcal{X}_l} \|X - \tilde{X}\|. \tag{3.4.1}$$

PROOF. The set $\mathcal{X}_l$ of $n \times k$ matrices $\tilde{X}$ of rank $\leqq l$ is defined by the condition that all minors of order $l+1$ of such matrices are equal to zero. Since these minors are polynomials in the elements of the matrices $\tilde{X}$, these equations define a closed set in the nk-dimensional space of matrices X. Let B be the ball of radius $\|X\|$ in this space, centered at X. Then $B \cap \mathcal{X}_l$ is compact, and is nonempty since it contains at least the zero matrix 0. Therefore the continuous function $f(\tilde{X}) = \|X - \tilde{X}\|$ has a minimum, $\hat{X}$, on $B \cap \mathcal{X}_l$, which is clearly the minimum on $\mathcal{X}_l$. (See Figure 3.4.2.) □

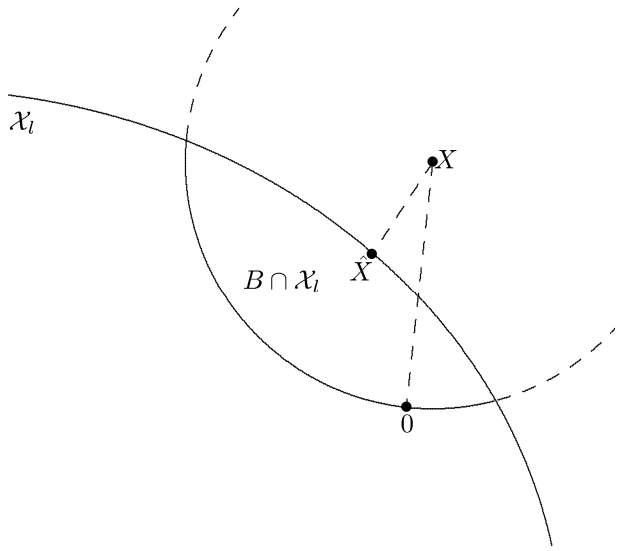

FIGURE 3.4.2

The following lemma and theorem have kindly been supplied by Joel Roberts of the School of Mathematics, University of Minnesota.

LEMMA 3.4.3 (ROBERTS). *Let E_{ij} be the $n \times k$ matrix with 1 in the i,jth position, and 0s elsewhere, and let A be any $n \times k$ matrix. Then there exists a real number $\lambda \neq 0$ such that the nk matrices*

$$\lambda E_{ij} - A \quad (i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, k) \tag{3.4.2}$$

form a basis in nk-dimensional space.

PROOF. If the set of matrices (3.4.2) is linearly dependent, then the $nk \times nk$ matrix whose columns are the successive columns of (3.4.2) is a matrix of the form $\lambda I_{nk} - M$, where M has all its nk columns equal to the column vector of columns of

A, and λ is an eigenvalue of M. But the eigenvalues of M are 0 with a multiplicity of $nk-1$, and just one other real number (namely, the sum of the elements of A). For any real number other than this one or 0, the set (3.4.2) is therefore linearly independent. □

THEOREM 3.4.1 (ROBERTS). *Let X be an $n \times k$ matrix of rank $> l$, and let $\hat{X} \in \mathcal{X}$ be a matrix of rank $\leqq l$ which is closest to X (in the Frobenius norm) among all matrices in the set $\mathcal{X}_l$ of $n \times k$ matrices of rank $\leqq l$. Then rank $\hat{X} = l$.*

PROOF. Suppose by way of contradiction that $\text{rank}(\hat{X}) < l$, and let λ be such as to satisfy Lemma 3.4.3. Since $\text{rank}(\lambda E_{ij}) = 1$ for $\lambda \neq 0$, we have, since multiplication of a matrix by a non-zero scalar does not affect its rank,

$$\text{rank}[(1-t)\hat{X} + t\lambda E_{ij}] \leqq \text{rank}(\hat{X}) + \text{rank}(\lambda E_{ij}) \leqq l \tag{3.4.3}$$

for any real number t, since the rank of the sum of two matrices is less than or equal to the sum of the ranks (because the column space of the sum of two matrices is contained in the sum of the column spaces of the two matrices, as is easily verified), and (3.4.3) holds for $t = 0$ and $t = 1$. Thus, of all points on the line $(1-t)\hat{X}+t\lambda E_{ij}$, the closest to X will be $\hat{X}$, since $\hat{X}$ has been assumed to be a closest point to X of all $\tilde{X}$ with $\tilde{X} \in \mathcal{X}_l$. Thus, the matrix $X - \hat{X}$ is perpendicular to the matrix $\hat{X} - \lambda E_{ij}$, since the shortest distance from a point to a line is along the perpendicular. But then $X-\hat{X}$ is perpendicular to all of nk-dimensional space, since $\hat{X}-\lambda E_{ij}$ is a basis for this space, by Lemma 3.4.3. Thus, $X - \hat{X} = 0$, i.e., $X = \hat{X}$. This contradiction establishes that $\text{rank}(\hat{X}) = l$. (See Figure 3.4.3.) □

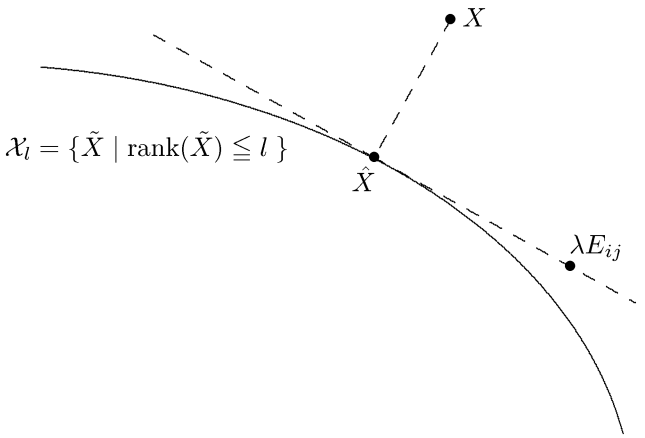

FIGURE 3.4.3

The Eckart-Young theorem states that any $n \times k$ matrix X of rank $p \leqq m = \min(n, k)$ can be best approximated in terms of the Frobenius norm by an $n \times k$ matrix of rank $l < p$, and that when the singular values of X are arranged in descending order, this approximating matrix can be obtained by striking out the singular values $s_{l+1}, s_{l+2}, \ldots, s_p$ of X. The theorem goes back to Eckart & Young

(1936, 1939), who tacitly assumed the first part to be true, and for the second part provided only a heuristic proof. This was followed by a more complete proof by Householder & Young (1938) and a much more detailed but still somewhat incomplete proof by Stewart (1973, pp. 322–3) (cf. Chipman 1997). A number of incorrect proofs have also appeared.[1] It was subsequently pointed out by Stewart & Sun (1990, pp. 208–210) that the theorem had already been proved (for integral operators and the Hilbert-Schmidt norm) by Schmidt (1907), and subsequently by Mirsky (1960) for unitarily invariant norms. Thus they refer to it as the "Schmidt-Mirsky theorem". Mirsky's proof (1960, Theorem 3) relied on the apparatus of symmetric gauge functions introduced by von Neumann (1937). For further discussion see Meyer (1993).

The following proof is based on the method followed in Stewart (1973).

THEOREM 3.4.2 (ECKART-YOUNG, SCHMIDT, MIRSKY). *Let X be a given $n \times k$ matrix of rank $p > l$, where $p \leqq m = \min(n, k)$. Then a matrix $\tilde{X}$ that minimizes $\|X - \tilde{X}\|$ over the set $\mathcal{X}_l = \{\tilde{X} \mid \mathrm{rank}(\tilde{X}) \leqq l\}$ is given by*

$$\hat{X} = QD_{(l)}P', \tag{3.4.4}$$

where

$$X = QDP' \tag{3.4.5}$$

is a singular-value decomposition of X, and the $n \times k$ matrix $D_{(l)}$ is obtained from D by replacing all but a set of its l largest diagonal elements by 0s. Further, $\hat{X}$ is a minimizer if and only if it is obtained in this way.

PROOF. By Lemma 3.4.2 such a matrix $\hat{X}$ exists, and by Theorem 3.4.1 it has rank l.

Let a singular-value decomposition of $\hat{X}$ be denoted

$$\hat{X} = \hat{Q}\hat{D}\hat{P}', \tag{3.4.6}$$

where D is an $n \times k$ diagonal matrix of the form

$$\hat{D} = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}, \tag{3.4.7}$$

where in turn S is an $l \times l$ diagonal matrix $\mathrm{diag}(s_1, s_2, \ldots, s_l)$ with $s_1 \geqq s_2 \geqq \ldots \geqq s_l > 0$. The main task of the proof is to show that $\hat{D} = D_{(l)}$, where the latter is obtained from D in the manner described in the statement of the theorem.

Define

$$\bar{D} = \hat{Q}'X\hat{P} \tag{3.4.8}$$

and partition it conformably with $\hat{D}$, as

$$\bar{D} = \begin{bmatrix} \bar{D}_{11} & \bar{D}_{12} \\ \bar{D}_{21} & \bar{D}_{22} \end{bmatrix}, \tag{3.4.9}$$

[1] Cf. Golub & Kahan (1965, p. 220); Rao (1965, p. 56; 1973, p. 70); Ben-Israel & Greville (1974, pp. 246–9; 1980, pp. 246–9). These are discussed in Chipman (1997).

$\bar{D}_{11}$ being of order $l \times l$ and $\bar{D}_{22}$ of order $(n-l) \times (k-l)$. $\bar{D}$ has rank p. Owing to the orthogonal invariance of the Frobenius norm, it follows from (3.4.6) and (3.4.8) that (3.4.1) is equivalent to

$$\|\bar{D} - \hat{D}\| = \min_{\text{rank}(\tilde{D}) \le l} \|\bar{D} - \tilde{D}\|. \tag{3.4.10}$$

We now show in Steps 1–3 below that the matrix $\bar{D}$ of (3.4.9) must be of the form

$$\bar{D} = \begin{bmatrix} S & 0 \\ 0 & \bar{D}_{22} \end{bmatrix}. \tag{3.4.11}$$

In Step 4 we will show that $\bar{D}$ is orthogonally equivalent to a diagonal matrix, D.

Step 1. First we show that $\bar{D}_{12} = 0$. Suppose not. Then the matrix

$$\tilde{D} = \begin{bmatrix} S & \bar{D}_{12} \\ 0 & 0 \end{bmatrix}$$

has the same rank as S, which is l, and

$$\|\bar{D} - \tilde{D}\| = \left\| \begin{bmatrix} \bar{D}_{11} - S & 0 \\ \bar{D}_{21} & \bar{D}_{22} \end{bmatrix} \right\| < \left\| \begin{bmatrix} \bar{D}_{11} - S & \bar{D}_{12} \\ \bar{D}_{21} & \bar{D}_{22} \end{bmatrix} \right\| = \|\bar{D} - \hat{D}\|.$$

Now define $\tilde{X} = \hat{Q}\tilde{D}\hat{P}'$; this matrix also has rank l. Then by the orthogonal invariance of the Frobenius norm we have

$$\|X - \tilde{X}\| = \|D - \tilde{D}\| < \|D - \hat{D}\| = \|X - \hat{X}\|. \tag{3.4.12}$$

Therefore $\tilde{X}$, which has rank l, is closer to X than $\hat{X}$. But this contradicts the hypothesis that $\hat{X}$ is a closest matrix to X among all $n \times k$ matrices $\tilde{X}$ of rank l. Therefore $\bar{D}_{12} = 0$.

Step 2. That $\bar{D}_{21} = 0$ is proved in similar fashion.

Step 3a. Now we show that $\text{rank}(\bar{D}_{11}) = l$. Suppose not; then since $\text{rank}(\bar{D}) = p > l$, and $\bar{D}_{12}$ and $\bar{D}_{21}$ have been shown to be zero, we can find a partition

$$\bar{D}_{22} = \begin{bmatrix} \bar{D}_{22,11} & \bar{D}_{22,12} \\ \bar{D}_{22,21} & \bar{D}_{22,22} \end{bmatrix}$$

of the $(n-l) \times (n-k)$ matrix $\bar{D}_{22}$ (if necessary by temporarily interchanging its rows and columns) such that $\bar{D}_{22,11}$ is of order $(p-l) \times (p-l)$ and rank $p-l$, so that the $n \times k$ matrix

$$\tilde{D} = \begin{bmatrix} \bar{D}_{11} & 0 & 0 \\ 0 & \bar{D}_{22,11} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

has rank l. Accordingly,

$$\begin{aligned} \|\bar{D} - \tilde{D}\| &= \left\| \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \bar{D}_{22,12} \\ 0 & \bar{D}_{22,21} & \bar{D}_{22,22} \end{bmatrix} \right\| \\ &< \left\| \begin{bmatrix} \bar{D}_{11} - S & 0 & 0 \\ 0 & \bar{D}_{22,11} & \bar{D}_{22,12} \\ 0 & \bar{D}_{22,21} & \bar{D}_{22,22} \end{bmatrix} \right\| = \|\bar{D} - \hat{D}\|. \end{aligned}$$

As before, we define $\tilde{X} = \hat{Q}\tilde{D}\hat{P}'$; this matrix also has rank l. Then with this new $\tilde{X}$, (3.4.12) holds as before and we arrive at a contradiction. Therefore $\text{rank}(\bar{D}_{11}) = l$.

Step 3b. Next we show that $\bar{D}_{11} = S$. Suppose not. Then define

$$\tilde{D} = \begin{bmatrix} \bar{D}_{11} & 0 \\ 0 & 0 \end{bmatrix}.$$

As just shown, this matrix has rank l, and

$$\|\bar{D} - \tilde{D}\| = \left\| \begin{bmatrix} 0 & 0 \\ 0 & \bar{D}_{22} \end{bmatrix} \right\| < \left\| \begin{bmatrix} \bar{D}_{11} - S & 0 \\ 0 & \bar{D}_{22} \end{bmatrix} \right\| = \|\bar{D} - \hat{D}\|,$$

leading to a contradiction, as before. Therefore $\bar{D}_{11} = S$.

From Steps 1–3 it follows that $\bar{D}$ must be of the form (3.4.11).

Step 4. Now let

$$\bar{D}_{22} = Q_{22} R P_{22}' \tag{3.4.13}$$

be a singular-value decomposition of the $(n-l) \times (k-l)$ matrix $\bar{D}_{22}$, where Q_{22} and P_{22} are, respectively, $(n-l) \times (n-l)$ and $(k-l) \times (k-l)$ orthogonal matrices, and R is an $(n-l) \times (k-l)$ diagonal matrix of singular values of $\bar{D}_{22}$. Define further the partitions $\hat{P} = [\hat{P}_1, \hat{P}_2]$ and $\hat{Q} = [\hat{Q}_1, \hat{Q}_2]$ of $\hat{P}$ and $\hat{Q}$ into their first l and last $k-l$ and $n-l$ columns, respectively. Finally, define the rectangular $n \times k$ diagonal matrix

$$D = \begin{bmatrix} S & 0 \\ 0 & R \end{bmatrix} \tag{3.4.14}$$

and the $k \times k$ and $n \times n$ matrices

$$P = [\hat{P}_1, \hat{P}_2] \begin{bmatrix} I_l & 0 \\ 0 & P_{22} \end{bmatrix}, \quad Q = [\hat{Q}_1, \hat{Q}_2] \begin{bmatrix} I_l & 0 \\ 0 & Q_{22} \end{bmatrix}, \tag{3.4.15}$$

which are readily verified to be orthogonal. Then we verify from (3.4.15), (3.4.14), (3.4.13), (3.4.10), and (3.4.8) that

$$QDP' = \hat{Q}\bar{D}\hat{P}' = X; \tag{3.4.16}$$

thus, QDP' is a singular-value decomposition of X, in accordance with (3.4.5); and from the orthogonal invariance of the Frobenius norm, D is orthogonally equivalent to $\bar{D}$ (and to X):

$$\|D\| = \|\bar{D}\| = \|X\|. \tag{3.4.17}$$

On the other hand, it is clear from (3.4.15), (3.4.7), and (3.4.6) that

$$Q\hat{D}P' = \hat{Q}\hat{D}\hat{P}' = \hat{X}, \tag{3.4.18}$$

so that $Q\hat{D}P'$ is a singular-value decomposition of $\hat{X}$. It remains to show that $\hat{D} = D_{(r)}$, establishing (3.4.4).

From (3.4.16), (3.4.14), and (3.4.13) we have

$$\|X\|^2 = \|S\|^2 + \|R\|^2 = \|S\|^2 + \|\bar{D}_{22}\|^2, \tag{3.4.19}$$

so that X has the diagonal elements of S as l of its singular values, and the sum of squares of its remaining $m-l$ singular values is equal to $\|\bar{D}_{22}\|^2$. From (3.4.16), (3.4.18), (3.4.14), (3.4.7), and (3.4.13) we have

$$\|X - \hat{X}\| = \|D - \hat{D}\| = \|\bar{D}_{22}\| = \|R\|. \tag{3.4.20}$$

Since by hypothesis, (3.4.20) is a minimum (satisfying (3.4.1)), this can only be the case if, in (3.4.19), the diagonal elements of S are the l largest singular values of X, and those of R are the $m-l$ smallest (with possible ties). It follows that, if the singular values of X are ordered as $s_1 \geqq s_2 \geqq \ldots \geqq s_l \geqq s_{l+1} \geqq \ldots \geqq s_m$, S must contain $s_1, s_2, \ldots, s_l$, and R must contain $s_{l+1}, \ldots, s_m$. (If $s_l = s_{l+1}$, $\hat{X}$ is not unique.) Applying this requirement to (3.4.7) and (3.4.14) we have $\hat{D} = D_{(r)}$ and the main part of the theorem is proved.

We finally come to the last statement of the theorem. Let $X = \check{Q}\check{D}\check{P}'$ be any other singular-value decomposition of X, and let $\check{D}_{(r)}$ be obtained from $\check{D}$ by replacing all but a set of its r largest singular values by 0s. Define $\check{X} = \check{Q}\check{D}_{(r)}\check{P}'$. Then by the orthogonal invariance of the Frobenius norm we have

$$\|X - \check{X}\| = \|\check{D} - \check{D}_{(r)}\| = \|D - D_{(r)}\| = \|X - \hat{X}\|. \quad \square$$

The following elementary proof of Theorem 3.4.2 is based on that of Neudecker in Chipman (1997, pp. 80–81) and Magnus & Neudecker (1999, pp. 359–361). First, a simple lemma will be used.

LEMMA 3.4.4. *Let A, B be $n \times n$ matrices, where B is symmetric. Then $tr(AB) = tr(A'B)$.*

PROOF. From the symmetry of B and Lemma 1.1.1,

$$\mathrm{tr}(AB) = \mathrm{tr}(AB)' = \mathrm{tr}(B'A') = \mathrm{tr}(BA') = \mathrm{tr}(A'B). \quad \square$$

ALTERNATIVE PROOF OF THEOREM 3.4.2 (NEUDECKER).[2] Let $\hat{X} \in \mathcal{X}_l$ be closest to the given $n \times k$ matrix X in the Frobenius norm. (The existence of $\hat{X}$ is assured by Lemma 3.4.2.) Its rows may be expressed without loss of generality as linear combinations of l $1 \times k$ orthonormal vectors, i.e.,

$$\hat{X} = AB', \quad B'B = I_l$$

where A is $n \times l$ and B is $k \times l$. By Theorem 3.4.1, $\hat{X}$ has rank l; therefore A must have rank l. We wish to find A and B that solve the problem

$$\underset{A,B}{\text{Minimize}}\ \mathrm{tr}[(X - AB')'(X - AB')] \quad \text{subject to} \quad B'B = I,$$

[2]This proof, like the previous one, assumes the truth of Lemma 3.4.2 and Theorem 3.4.1.

or equivalently,

$$\underset{A,B}{\text{Maximize}}\ \psi \equiv 2\ \text{tr}(BA'X) - \text{tr}(A'A) \quad \text{subject to} \quad B'B = I.$$

Setting up the Lagrangean expression

$$\varphi = 2\ \text{tr}(BA'X) - \text{tr}(A'A) - \text{tr}[L(B'B - I)],$$

we note that since $B'B$ is symmetric, from Lemma 3.4.4 we may without loss of generality assume the Lagrangean multiplier matrix L to be symmetric (or replaced by $\frac{1}{2}(L + L')$). Using this symmetry we obtain for variations in A and B

$$\begin{aligned} &d\varphi \\ &= 2\ \text{tr}(B'X')dA + 2\ \text{tr}(A'X)dB - 2\ \text{tr}(A')dA - \text{tr}(LB')dB - \text{tr}(L'B')dB \\ &= 2\ \text{tr}(B'X' - A')dA + 2\ \text{tr}(A'X - LB')dB. \end{aligned}$$

Setting $d\varphi = 0$ for arbitrary dA and dB yields, with the given constraint,

(i) $$XB = A$$

(ii) $$A'X = LB'$$

(iii) $$B'B = I.$$

From these three equations we obtain

(iv) $$A'\underline{A} \underset{\text{(i)}}{=} \underline{A'XB} \underset{\text{(ii)}}{=} L\underline{B'B} \underset{\text{(iii)}}{=} L.$$

Since $\text{rank}(A) = l$, $L = A'A$ is positive-definite. From equations (i) and (ii) and the symmetry of L we obtain

(v) $$X'\underline{XB} \underset{\text{(i)}}{=} \underline{X'A'} \underset{\text{(ii)}}{=} BL' = BL.$$

From (ii) and (iv) it follows that

$$\psi = 2\ \text{tr}(BA'X) - \text{tr}(A'A) = 2\ \text{tr}(BLB') - \text{tr}(L) = \text{tr}(L),$$

which is to be a maximum.

Note from (i) that

$$\hat{X} = AB' = XBB',$$

whch states (since BB' is idempotent and symmetric) that $\hat{X}$ is a perpendicular projection of the rows of X onto an l-dimenional subspace.

Let T be an orthogonal matrix such that

$$L = T\Lambda T'$$

where Λ is diagonal, and define

$$\tilde{A} = AT, \quad \tilde{B} = BT.$$

Then

$$\tilde{A}'\tilde{A} = T'A'AT = T'LT = \Lambda$$

and

$$\tilde{B}'\tilde{B} = T'B'BT = T'T = I.$$

Equations (i) to (iii) above then become

(i′) $$X\tilde{B} = \tilde{A}$$

(ii′) $$\tilde{A}'X = T'LB' = T'LT\tilde{B}' = \Lambda\tilde{B}'$$

(iii′) $$\tilde{B}'\tilde{B} = I.$$

From these equations it follows that

$$X'X\tilde{B} = \tilde{B}\Lambda \quad \text{and} \quad \tilde{B}'\tilde{B} = I.$$

Thus, Λ, whose trace is to be maximized (being equal to the trace of L), is a diagonal matrix of l eigenvalues of $X'X$, and $\tilde{B}$ is the matrix whose l columns constitute an associated orthonormal set of l eigenvectors of $X'X$. $\operatorname{tr}(\Lambda)$ is maximized when these l eigenvalues are a set of l largest eigenvalues of $X'X$.

Now let X have the singular-value decomposition (3.4.4); its singular values s_i (the diagonal elements of D) are the positive square roots of the eigenvalues λ_i of Λ (from $X'X = PD'DP'$), so since $\hat{X}$ has been assumed to be the closest to X, its singular values must be the l largest singular values of D. □

3.5 Reduced-rank estimation

It was proposed by Marquardt (1970)—with respect to the regression model

(3.5.1) $$y = X\beta + \varepsilon; \quad E\{\varepsilon\} = 0; \quad \operatorname{Var}\{\varepsilon\} = \sigma^2\Omega,$$

in the special case $\Omega = I$—that when the obervation matrix X is ill-conditioned (as defined by its condition number—see Definition 1.3.3), one can obtain an estimator of β in this regression model which has lower scalar mean-square error than that of the least-squares estimator, by finding the best approximation to X by a matrix $X_{(l)}$ of rank $l < k$ and then replacing the least-squares estimator $\tilde{\beta} = X^\dagger y$ by the estimator $\hat{\beta}_{(l)} = X^\dagger_{(l)} y$. The theory behind this procedure—and a generalization—will be developed in this section.

LEMMA 3.5.1. *Let the $n \times k$ matrix X, of rank $k < n$, have singular-value decomposition $X = Q_1 S P'$ (as in (3.2.8)), where the singular values of X are arranged in descending order $s_1 \geqq s_2 \geqq \ldots \geqq s_k > 0$, and let its best approximation by an $n \times k$ matrix of rank l, in terms of the Frobenius norm of Definition 3.4.1, be given by $X_{(l)} = Q_1 S_{(l)} P'$, where $S_{(l)}$ is obtained from S by replacing s_i by zero for $i = l+1, \ldots, k$. Then*

(3.5.2) $$X_{(l)} = XP_1P_1' = X[I - P_2P_2'],$$

where $P = [P_1, P_2]$ is a partition of the $k \times k$ orthogonal matrix P into its first l and last $r = k - l$ columns. Further, the Moore-Penrose generalized inverse of $X_{(l)}$ is given by

(3.5.3) $$X^\dagger_{(l)} = P_1P_1'X^\dagger = [I - P_2P_2']X^\dagger.$$

PROOF. We denote the partitioned matrices

$$S = \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \quad \text{and} \quad S_{(l)} = \begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{3.5.4}$$

where S_1 and S_2 are diagonal matrices of orders $l \times l$ and $r \times r$ respectively. Then

$$X_{(l)} = Q_1 \begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} P_1' \\ P_2' \end{bmatrix} = Q_1 \begin{bmatrix} S_1 P_1' \\ 0 \end{bmatrix}$$

while

$$XP_1P_1' = Q_1 \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} P_1' \\ P_2' \end{bmatrix} P_1P_1' = Q_1 \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} P_1' \\ 0 \end{bmatrix} = Q_1 \begin{bmatrix} S_1P_1' \\ 0 \end{bmatrix},$$

and these are the same, establishing (3.5.2). Likewise,

$$X_{(l)}^{\dagger} = PS_{(l)}^{\dagger}Q_1' = [P_1 \quad P_2] \begin{bmatrix} S_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q_1' = [P_1S_1^{-1} \quad 0]Q_1',$$

while

$$\begin{aligned} P_1P_1'X^{\dagger} &= P_1P_1'[P_1 \quad P_2] \begin{bmatrix} S_1^{-1} & 0 \\ 0 & S_2^{-1} \end{bmatrix} Q_1' \\ &= [P_1 \quad 0] \begin{bmatrix} S_1^{-1} & 0 \\ 0 & S_2^{-1} \end{bmatrix} Q_1' = [P_1S_1^{-1} \quad 0]Q_1', \end{aligned}$$

and these are the same, establishing (3.5.3). □

In order to pursue Marquardt's result in the general case $\text{Var}\{\varepsilon\} = \sigma^2\Omega$, and in terms of the matrix concept of mean-square error, it is clear that we will want to employ an estimator that is an oblique generalized inverse of a reduced-rank matrix with respect to Ω. This may be accomplished by premultiplying the variables y, X, and ε in (3.5.1) by $\Omega^{-1/2}$, where $\Omega^{1/2}$ is a symmetric positive-definite square root of Ω, and defining

$$\dot{y} = \Omega^{-1/2}y, \quad \dot{X} = \Omega^{-1/2}X, \quad \dot{\varepsilon} = \Omega^{-1/2}\varepsilon. \tag{3.5.6}$$

We see that

$$\text{Var}\{\dot{\varepsilon}\} = \sigma^2 I.$$

Defining

$$X^{\ddagger} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}, \tag{3.5.7}$$

we verify that

$$X^{\ddagger} = \dot{X}^{\dagger}\Omega^{-1/2}. \tag{3.5.8}$$

This transformation of variables amounts to replacing the Frobenius norm $\|X\|$ of Definition 3.4.1 by the Ω-norm defined by

$$\|X\|_\Omega = \sqrt{\mathrm{tr}(X'\Omega^{-1}X)}. \tag{3.5.9}$$

In terms of this norm the "best approximation" of the matrix $X = \Omega^{1/2}\dot{X}$ by an $n \times k$ matrix of rank l is then

$$X_{[l]} = \Omega^{1/2}\dot{X}_{(l)} = \Omega^{1/2}\dot{X}\dot{P}_1\dot{P}_1' = X\dot{P}_1\dot{P}_1' = X(I - \dot{P}_2\dot{P}_2'), \tag{3.5.10}$$

where

$$\dot{X} = \dot{Q}_1\dot{S}\dot{P}' \tag{3.5.11}$$

is a singular-value decomposition of $\dot{X}$ and $\dot{P} = [\dot{P}_1, \dot{P}_2]$ is a partition of the orthogonal matrix $\dot{P}$ into its first l and last $r = k - l$ columns. Then, defining

$$X_{[l]}^\ddagger = \dot{X}_{(l)}^\dagger\Omega^{-1/2} = \dot{P}_1\dot{P}_1'\dot{X}^\dagger\Omega^{-1/2} = \dot{P}_1\dot{P}_1'X^\ddagger = (I - \dot{P}_2\dot{P}_2')X^\ddagger, \tag{3.5.12}$$

we verify that $X_{[l]}^\ddagger$ is the oblique generalized inverse of $X_{[l]}$ with respect to $U = I$ and $V = \Omega$.

The following generalizes Marquardt's theorem (1970, pp. 591–612):

THEOREM 3.5.1. *Let $X_{[l]}$ be the best approximation of X by a matrix of rank $l < k$ in terms of the Ω-norm (3.5.9), and let $X_{[l]}^\ddagger$ be defined by (3.5.12). Then:*

(a) a necessary and sufficient condition for the reduced-rank estimator

$$\hat{\beta}_{[l]} = X_{[l]}^\ddagger y = (I - \dot{P}_2\dot{P}_2')X^\ddagger y \tag{3.5.13}$$

to have matrix mean-square error no greater than that of the generalized least-squares (Gauss-Markov) estimator

$$\tilde{\beta} = X^\ddagger y \tag{3.5.14}$$

is that

$$\sigma^{-2}\dot{P}_2'\beta\beta'\dot{P}_2 \preccurlyeq \dot{S}_2^{-2}. \tag{3.5.15}$$

(b) A sufficient condition for (3.5.15) to hold is that

$$\lambda \equiv \frac{\beta'\dot{P}_2\dot{S}_2^2\dot{P}_2'\beta}{\sigma^2} \leqq 1. \tag{3.5.16}$$

(c) If $r = k - l > 1$, condition (3.5.16) is sufficient for the inequality (3.5.15) to be strict, and if $r = 1$ then conditions (3.5.15) and (3.5.16) are equivalent, hence both necessary and sufficient.

PROOF. (a) From the singular-value decomposition of X we readily compute the mean-square error of the generalized least-squares estimator,

$$\mathrm{Risk}\{\tilde{\beta}\} = \sigma^2\dot{P}\dot{S}^{-2}\dot{P}' = \sigma^2(\dot{P}_1\dot{S}_1^{-2}\dot{P}_1' + \dot{P}_2\dot{S}_2^{-2}\dot{P}_2'). \tag{3.5.17}$$

To compute that of the reduced-rank estimator we observe that

$$\begin{aligned}\hat{\beta}_{[l]} - \beta &= X^{\ddagger}_{[l]}\varepsilon - (I - \dot{P}_1\dot{P}_1')\beta \\ &= \dot{P}_1\dot{P}_1'X^{\ddagger}\varepsilon - \dot{P}_2\dot{P}_2'\beta.\end{aligned}$$

Consequently, from the singular-value decomposition (3.5.11) we have

$$\begin{aligned}\operatorname{Risk}\{\hat{\beta}_{[l]}\} &= \sigma^2\dot{P}_1\dot{P}_1'(\dot{X}'\dot{X})^{-1}\dot{P}_1\dot{P}_1' + \dot{P}_2\dot{P}_2'\beta\beta'\dot{P}_2\dot{P}_2' \\ &= \sigma^2\dot{P}_1\dot{P}_1'[\dot{P}_1, \dot{P}_2]\begin{bmatrix}\dot{S}_1^{-2} & 0\\ 0 & \dot{S}_2^{-2}\end{bmatrix}\begin{bmatrix}\dot{P}_1'\\ \dot{P}_2'\end{bmatrix}\dot{P}_1\dot{P}_1' + \dot{P}_2\dot{P}_2'\beta\beta'\dot{P}_2\dot{P}_2' \\ &= \sigma^2\dot{P}_1\dot{S}_1^{-2}\dot{P}_1' + \dot{P}_2\dot{P}_2'\beta\beta'\dot{P}_2\dot{P}_2'.\end{aligned} \tag{3.5.18}$$

The difference between the two mean-square errors (3.5.17) and (3.5.18) is therefore

$$\operatorname{Risk}\{\tilde{\beta}\} - \operatorname{Risk}\{\hat{\beta}_{[l]}\} = \sigma^2\dot{P}_2[\dot{S}_2^{-2} - \sigma^{-2}\dot{P}_2'\beta\beta'\dot{P}_2]\dot{P}_2'. \tag{3.5.19}$$

This is nonnegative-definite if and only if (3.5.15) holds.

(b) The sufficient condition (3.5.16) is obtained by making use of the generalized matrix Cauchy-Schwarz inequality $AVA' \succcurlyeq AX(X'V^{-1}X)^{-1}X'A'$ (Lemma 2.4.1) with I_r substituted for "A", $\dot{S}_2^{-2}$ for "V", and $\dot{P}_2'\beta$ for "X", yielding the first inequality in

$$\dot{S}_2^{-2} \succcurlyeq \dot{P}_2'\beta(\beta'\dot{P}_2\dot{S}_2^2\dot{P}_2'\beta)^{-1}\beta'\dot{P}_2 = \lambda^{-1}\sigma^{-2}\dot{P}_2'\beta\beta'\dot{P}_2 \succcurlyeq \sigma^{-2}\dot{P}_2'\beta\beta'\dot{P}_2. \tag{3.5.20}$$

The equality in (3.5.20) follows from the definition of λ in (3.5.16), and the second inequality in (3.5.20) follows from $\lambda^{-1} \geqq 1$. The entire inequality in (3.5.20) is simply (3.5.15).

(c) Since the matrix $\dot{S}_2^{-2}$ on the right side of the inequality (3.5.15) has rank r, whereas the matrix $\beta\beta'$ on the left has rank 1, it is impossible for equality to hold in (3.5.15) unless $r = 1$. Thus, the inequality is necessarily strict if $r > 1$. If $r = 1$, since $\dot{P}_2$ is $k \times r$ and $\dot{S}_2$ is $r \times r$, both sides of (3.5.15) are scalars, hence (3.5.15) reduces to (3.5.16), which is then necessary and sufficient. □

REMARK 3.5.1. *It will be shown in the next chapter (Theorem 4.2.3 and formula (4.2.30)) that the reduced-rank estimator (3.5.13) is the minimum-variance conditionally unbiased estimator of β subject to the homogeneous linear restriction $\dot{P}_2'\beta = 0$.*

The duality referred to in Remark 3.5.1 was apparently first noticed in an unpublished paper by Johnson & Wallace (1969).

COROLLARY 3.5.1. *A sufficient condition for* $\operatorname{Risk}\{\hat{\beta}_{[l]}\} \preccurlyeq \operatorname{Risk}\{\tilde{\beta}\}$ *is*

$$\frac{\beta'\dot{P}_2\dot{P}_2'\beta}{\sigma^2} \leqq \frac{1}{\dot{s}_{l+1}^2}. \tag{3.5.21}$$

A sufficient condition for (3.5.21) is in turn

$$\frac{\beta'\beta}{\sigma^2} \leqq \frac{1}{\dot{s}_{l+1}^2}, \tag{3.5.22}$$

where the $\dot{s}_i^2$ are the eigenvalues of $X'\Omega^{-1}X$, in descending order.

PROOF. Applying the matrix Cauchy-Schwarz inequality (Lemma 2.4.1) for $A = V = I$, we have $I \succcurlyeq X(X'X)^{-1}X'$, hence replacing "$X$" by $\dot{P}_2'\beta$ we obtain

$$I_r \succcurlyeq \dot{P}_2'\beta(\beta'\dot{P}_2\dot{P}_2'\beta)^{-1}\beta'\dot{P}_2, \quad \text{or} \quad (\beta'\dot{P}_2\dot{P}_2'\beta)I_r \succcurlyeq \dot{P}_2'\beta\beta'\dot{P}_2, \tag{3.5.23}$$

Assuming (3.5.21) to hold it follows that

$$\dot{S}_2^{-2} \succcurlyeq \dot{s}_{l+1}^{-2} I_r \succcurlyeq \sigma^{-2}(\beta'\dot{P}_2\dot{P}_2'\beta)I_r \succcurlyeq \sigma^{-2}\dot{P}_2'\beta\beta'\dot{P}_2,$$

i.e., (3.5.15) holds. Since $I - \dot{P}_2\dot{P}_2' = \dot{P}_1\dot{P}_1' \succcurlyeq 0$, $\beta'\dot{P}_2\dot{P}_2'\beta \leqq \beta'\beta$, hence (3.5.21) follows from (3.5.22). □

Thus, if one is able to impose an *a priori* upper bound on $\beta'\beta/\sigma^2$, and is one has data on the singular values $\dot{s}_i$ of $\dot{X} = \Omega^{-1/2}X$, one may select the appropriate dimensionality l by choosing the largest singular value that satisfies (3.5.22)—or still more accurately, (3.5.21).

We may note that condition (3.5.21) is stronger than, since it implies, (3.5.16); for $\dot{S}_2^2 \preccurlyeq \dot{s}_{l+1}^2 I_r$ from the definition of $\dot{S}_2^2$, hence (3.5.21) implies

$$\lambda = \frac{\beta'\dot{P}_2\dot{S}_2^2\dot{P}_2'\beta}{\sigma^2} \leqq \dot{s}_{l+1}^2 \frac{\beta'\dot{P}_2\dot{P}_2'\beta}{\sigma^2} \leqq 1.$$

While the still stronger condition (3.5.22) has the advantage of not requiring computation of $\dot{P}_2$, condition (3.5.16) has the advantage—as will become clear in the next chapter—that it can be tested from the data (provided of course ε is normally distributed), since λ is the noncentrality parameter of the noncentral F-distribution used for testing the null hypothesis $\lambda = 1$ ($\dot{P}_2'\beta = 0$) against the alternative hypothesis $\lambda > 1$ ($\dot{P}_2'\beta \neq 0$); see Remark 3.5.1 above and Theorem 4.3.1 below.

Let us now consider Marquardt's criterion. Marquardt used a scalar criterion of mean-square error, equal to the trace of the matrix measure used here. If an estimator $\hat{\beta}_{[l]}$ has lower matrix-mean-square error than $\tilde{\beta}$, this means that *each component* of $\hat{\beta}_{[l]}$ has lower mean-square error than the corresponding component of $\tilde{\beta}$. With the scalar definition, this need not be the case; some components could have lower, others higher, mean-square error; only an average of them has lower mean-square error. It is to be expected, then, that the matrix definition requires more stringent conditions. Taking the trace of both sides of (3.5.15) we obtain

$$\frac{\beta'\dot{P}_2\dot{P}_2'\beta}{\sigma^2} \leqq \sum_{i=l+1}^{k} \frac{1}{\dot{s}_i^2}. \tag{3.5.24}$$

This is clearly a much weaker condition than (3.5.21); Marquardt (1970, p. 601) actually specified the somewhat more stringent condition

$$\frac{\beta'\beta}{\sigma^2} \leqq \sum_{i=l+1}^{k} \frac{1}{\dot{s}_i^2}, \tag{3.5.25}$$

which implies (3.5.24). Clearly, (3.5.21) and (3.5.22) imply (3.5.24) and (3.5.25) respectively. However, Marquardt's condition has a curious interpretation: if the inequality is satisfied for some $l < k$, then it is satisfied for any $l' < l$; in particular, if it is satisfied for $l = k - 1$, then it is also satisfied for $l = 1$. Thus it provides no guide for choosing the appropriate rank, l. On the other hand, (3.5.22), or better still (3.5.21), provides just such a guide.

3.6 Exercises

1. An investigator wishes to estimate the parameters β_1 and β_2 in the model

$$y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \varepsilon_t, \quad E\{\varepsilon_t\} = 0, \quad E\{\varepsilon_t\varepsilon_{t'}\} = \delta_{tt'}\sigma^2$$

$$\text{where } t, t' = 1, 2, \ldots, n \quad \text{and} \quad \delta_{tt'} = \begin{cases} 1 & \text{for } t = t', \\ 0 & \text{for } t \neq t', \end{cases}$$

where the column vectors $x^j = (x_{1j}, x_{2j}, \ldots, x_{nj})'$ have been normalized to have length 1 for $j = 1, 2$. This investigator notices that the correlation $r = \sum_{t=1}^{n} x_{t1}x_{t2}$ between the two independent variables is rather close to 1, and therefore decides to estimate the two parameters by finding the best approximation of $X = [x^1, x^2]$ by an $n \times 2$ matrix $X_{(1)}$ of rank 1 (the distance between two matrices X and X^* being defined by the Frobenius norm $\|Z\|$ of their difference $Z = X - X^*$), and then estimating the 2×1 vector β by $\hat{\beta}_{(1)} = X_{(1)}^{\dagger} y$ (the "Marquardt estimator"), where $\dagger$ denotes the Moore-Penrose generalized inverse.

(a) Show that the best approximation of X by a matrix of rank 1 is the $n \times 2$ matrix each of whose columns is the average of the two.

(b) Obtain the formula for the Marquardt estimator.

(c) Show that the Marquardt estimator estimates each parameter β_j by the simple average of the least-squares estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$.

(d) Find the expressions for the (scalar) mean-square errors

$$E\{(\hat{\beta}_{(1)} - \beta)'(\hat{\beta}_{(1)} - \beta)\} \quad \text{and} \quad E\{(\tilde{\beta} - \beta)'(\tilde{\beta} - \beta)\}$$

of $\hat{\beta}_{(1)}$ and $\tilde{\beta}$ respectively.

(e) Show that $\hat{\beta}_{(1)}$ has lower mean-square error than $\tilde{\beta}$ if and only if

$$\frac{(\beta_1 - \beta_2)^2}{2\sigma^2} < \frac{1}{1 - r}.$$

2. What is the condition number of the matrix X of the previous question?

3.7 References

Bellman, Richard. *Introduction to Matrix Analysis.* New York: McGraw-Hill Book Company, Inc., 1960. xx, 328 pp.

Ben Israel, Adi, and Greville, Thomas N. E. *Generalized Inverses: Theory and Applications.* New York: John Wiley & Sons, Inc., 1974. xi, [3], 395 pp. Reprint edition with corrections, Huntington, New York: Robert E. Krieger Publishing Company, 1980.

Chipman, John S. "On Least Squares with Insufficient Observations." *Journal of the American Statistical Association*, 59 (December 1964), 1078–1111. Corrigendum: 60 (December 1965), 1249.

Chipman, John S. "Estimation and Aggregation in Econometrics: An Application of the Theory of Generalized Inverses." In Nashed (1976), 549–769.

Chipman, John S. "'Proofs' and Proofs of the Eckart-Young Theorem," with an Appendix by Heinz Neudecker. In Jerome A. Goldstein, Neil E. Gretsky, and J. J. Uhl, Jr. (eds.), *Stochastic Processes and Functional Analysis. In Celebration of M. M. Rao's 65th Birthday* (New York: Marcel Dekker, Inc., 1997), 71–83.

Chipman, John S. "Linear Restrictions, Rank Reduction, and Biased Estimation in Linear Regression." *Linear Analysis and Its Applications*, 289 (1999), 55–74.

Courant, R. "Zur Theorie der kleinen Schwingungen." *Zeitschrift für angewandte Mathematik und Mechanik*, 2 (1922), 278–285.

Eckart, Carl, and Young, Gale. "The Approximation of One Matrix by Another of Lower Rank." *Psychometrika*, 1 (September 1936), 211–218.

Eckart, Carl, and Young, Gale. "A Principal Axis Transformation for Non-Hermitian Matrices." *Bulletin of the American Mathematical Society*, 45 (February 1939), 118–121.

Fischer, Ernst. "Über quadratische Formen mit reelen Koeffizienten." *Monatsheft für Mathematik und Physik*, 16 (1905), 234–249.

Frisch, Ragnar. *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo: Universitetets Økonomisk Institutt, Publikasjon nr. 5, 1934. 192 pp.

Golub, Gene H., and Kahan, W. "Calculating the Singular Values and Pseudo-Inverse of a Matrix." *SIAM Journal of Numerical Analysis* [B], 2 (1965), 205–224.

Golub, Gene H., and Reinsch, C. "Singular Value Decomposition and Least Squares Solutions." *Numerische Mathematik*, 14 (1970), 403–420. Reprinted in Wilkinson & Reinsch (1971), 134–151.

Golub, Gene H., and Styan, George P. H. "Numerical Computations for Univariate Linear Models." *Journal of Statistical Computation and Simulation*, 2 (1973), 253–274.

Golub, Gene H., and Van Loan, Charles F. *Matrix Computations*. Baltimore: The Johns Hopkins University Press, 1983. xvi, [2], 476, [1] pp.

Gunst, Richard F., and Mason, Robert L. "Generalized Mean Squared Error Properties of Regression Estimators." *Communications in Statistics*, A5 (1976), 1501–1508.

Gunst, Richard F., and Mason, Robert L. "Biased Estimation in Regression: An Evaluation using Mean Squared Error." *Journal of the American Statistical Association*, 72 (September 1977), 616–628.

Gunst, Richard F., Webster, John T., and Mason, Robert L. "A Comparison of Least Squares and Latent Root Regression Estimators." *Technometrics*, 18 (1976), 75–83.

Hawkins, Douglas M. "On the Investigation of Alternative Regressions by Principal Component Analysis." *Applied Statistics*, 22 (1973), 275–286.

Householder, A. S., and Young, Gale. "Matrix Approximations and Latent

Roots." *American Mathematical Monthly*, 45 (March 1938), 165–171.

Johnson, Thomas, and Wallace, T. D. "Principal Components and Multicollinearity." Workshop Discussion Paper, Department of Economics, North Carolina State University, Raleigh, NC, June 1969.

Lott, William F. "The Optimal Set of Principal Component Restrictions on a Least-Squares Regression." *Communications in Statistics*, 2 (1973), 449–463.

Magnus, Jan R., and Neudecker, Heinz. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Chichester and New York: John Wiley & Sons, 1991, 1994. Revised edition, c. 1999. xviii, 395, [8] pp.

Mansfield, Edward R., Webster, John T., and Gunst, Richard F. "An Analytic Variable Selection Technique for Principal Component Regression." *Applied Statistics*, 26 (1977), 34–40.

Marquardt, Donald W. "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation." *Technometrics*, 12 (August 1970), 591–612.

Mason, Robert L., Gunst, Richard F., and Webster, John T. "Regression Analysis and Problems of Multicollinearity." *Communications in Statistics*, 4 (1975), 277–292.

Massy, William F. "Principal Components Regression in Exploratory Statistical Research." *Journal of the American Statistical Association*, 60 (March 1965), 234–256.

Meyer, Renate. *Matrix-Approximation in der multivariaten Statistik.* Aachen: Verlag der Augustinus Buchhandlung, 1993. [6], 131, [2] pp.

Mirsky, L. "Symmetric Gauge Functions and Unitarily Invariant Norms." *Quarterly Journal of Mathematics*, Oxford Second Series, 11 (March 1960), 50–59.

Nashed, M. Zuhair (ed.). *Generalized Inverses and Applications.* New York: Academic Press, 1976. xiv, 1054 pp.

Neumann, J[ohn] von. "Some Matrix-Inequalities and Metrization of Matric-Space." *Tomsk University Review*, 1 (1937), 286–300.

Noble, B. "Methods for Computing the Moore-Penrose Generalized Inverse, and Related Matters." In Nashed (1976), 245–301.

Penrose, R. "A Generalized Inverse for Matrices." *Proceedings of the Cambridge Philosophical Society*, 51 (1955), 406–413.

Rao, C. Radhakrishna. *Linear Statistical Inference and Its Applications.* New York: John Wiley & Sons, 1965. xviii, [2], 522 pp. 2nd edition, 1973. xx, [2], 625 pp.

Rao, C. Radhakrishna. "Separation Theorems for Singular Values of Matrices and Their Applications in Multivariate Analysis." *Journal of Multivariate Analysis*, 9 (1979), 362–377.

Rao, C. Radhakrishna. "Matrix Approximations and Reduction of Dimen-

sionality in Multivariate Statistical Analysis." In Paruchuri R. Krishnaiah, ed., *Multivariate Analysis – V. Proceedings of the Fifth International Symposium on Multivariate Analysis* (Amsterdam: North-Holland Publishing Company, 1980), 3–22.

Schmidt, Erhard. "Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener." *Mathematische Annalen*, 63 (1907), 433–476.

Silvey, S. D. "Multicollinearity and Imprecise Estimation." *Journal of the Royal Statistical Society* [B], 31 (1969), 539–552.

Sondermann, Dieter. "Best Approximate Solutions to Matrix Equations under Rank Restrictions." Report No. 23/80, Institute for Advanced Studies, The Hebrew University of Jerusalem, Mount Scopus, Israel, August 1980.

Stewart, G. W. *Introduction to Matrix Computations*. New York: Academic Press, 1973. xi, [3], 441 pp.

Stewart, G. W., and Sun, Ji-guang. *Matrix Perturbation Theory*. San Diego: Academic Press, Inc., 1990. xvi, 365 pp.

Webster, John T., Gunst, Richard F., and Mason, Robert L. "Latent Root Regression Analysis." *Technometrics*, 16 (1974), 513–522.

Wilkinson, J. H., and Reinsch, C. H. (eds.). *Linear Algebra*. New York: Springer-Verlag, 1971. VIII, [2], 439, [3] pp.